PROTEIN INTERFACES IN CRYSTAL STRUCTURES: INSIGHTS FROM EVOLUTION

DISSERTATION

ZUR
ERLANGUNG DER NATURWISSENSCHAFTLICHEN DOKTORWÜRDE
(DR. SC. NAT.)

VORGELEGT DER
MATHEMATISCH-NATURWISSENCHAFTLICHEN FAKULTÄT
DER
UNIVERSITÄT ZÜRICH

VON
JOSE MANUEL DUARTE GAMERO
AUS
SPANIEN

PROMOTIONSKOMITEE
PROF. DR. AMEDEO CAFLISCH (VORSITZ)
PROF. DR. RAIMUND DUTZLER
DR. GUIDO CAPITANI (LEITUNG)

ZÜRICH, 2013

Table of Contents

Summary

This doctoral thesis revolves around understanding protein-protein interfaces as found in protein crystal structures solved via X-ray crystallography. Protein structures are known in atomic detail thanks to X-ray diffraction experiments performed nowadays mostly at synchrotron radiation sources. In these experiments, proteins are however not in their native solution environment. Instead, they are first crystallized so that they can strongly diffract the incoming X-ray beams and produce measurable diffraction patterns.

In the formation of the crystal an essential piece of information disappears: the interfaces that different polypeptide chains form among them in the solution environment are lost when the protein arranges into a crystal lattice. The diffraction data offers the detailed position of the atoms in the protein fold – the Tertiary Structure – but does not tell explicitly about the arrangement of the chains together into a Quaternary Structure. The crystal lattice thus contains two kinds of contacts among polypeptide chains: non-specific ones, consequence of the formation of the crystal lattice and specific ones that are biologically relevant.

We aim mainly at computationally distinguishing these two kinds of protein interfaces. The key difference between the two types is that biological interfaces have been subjected to the forces of evolution. Here, the abundant data coming from DNA sequencing technologies provides the required evolutionary background information that can be connected to the structural data. By combining the patterns of evolution seen in Multiple Sequence Alignments and the protein's 3-dimensional coordinates we try to detect the footprint of evolution on protein interfaces in order to differentiate them from crystal lattice contacts.

We show how we developed such a method and demonstrate that it is very effective at classifying biological interfaces from crystal contacts. Thus the method contributes greatly to the interpretation of protein crystal structures indicating the correct biological unit assembly that the proteins possess in their native environment.

The classification method was initially developed with soluble proteins in mind. In a second part of the study, however, we also proved its applicability to the interfaces found in crystals of membrane protein structures. A necessary step in this analysis was to compile a validated set of transmembrane protein-protein interfaces from the known set of membrane structures deposited in the Protein Data Bank. Such a dataset constitutes the first comprehensive compilation of validated transmembrane protein interfaces. Through it we have tried to establish the principles of how

interfaces assemble in the transmembrane region and how they compare to those of soluble proteins.

We thus established the applicability of the newly developed method, called EPPIC, in both the soluble protein and the membrane protein worlds. We finally offer a robust implementation of the method in a stand-alone software package and in a web graphical user interface, making it available to the wide structural biology and bioinformatics communities.

Zusammenfassung

Diese Arbeit befasst sich mit der Analyse von Protein-Protein-Kontaktoberflächen, wie sie bei der Proteinstrukturermittlung mittels Röntgenkristallographie auftreten. Die räumliche Struktur von Proteinen kann durch Röntgenbeugungsexperimente ermittelt werden. Dabei liegen die Proteine nicht wie in der biologischen Umgebung in gelöster Form vor, sondern werden zunächst kristallisiert, so dass die auftreffenden Röntgenstrahlen stark gebeugt werden, und zu messbaren Beugungsmustern führen.

Bei der Kristallbildung geht die Information verloren, welche spezifischen Kontakte die Proteine in der gelösten Umgebung bilden. Die Messdaten erlauben die Bestimmung der räumlichen Anordnung des gefalteten Proteins (die Tertiärstruktur) innerhalb der Kristallanordnung, nicht aber die Komplexbildung in der biologischen Umgebung (Quarternärstruktur). Anhand der Messdaten lässt sich also nicht unterscheiden zwischen Kontakten, die nur im Kristall auftreten und solchen, die in der biologischen Umgebung relevant sind.

Das Ziel dieser Arbeit ist, die Unterscheidung dieser beiden Arten von Kontakten mit Hilfe von computergestützten Methoden. Die Unterscheidung basiert auf der Tatsache, dass biologische Kontakte im Gegensatz zu kristallinen Kontakten evolutionärer Selektion unterworfen sind. Die moderne DNS-Sequenziertechnologie liefert eine große Menge an Daten über evolutionäre Prozesse in Biomolekülen. Durch die Kombination dieser Daten in der Form von multiplen Sequenzalignments mit den Proteinstrukturdaten detektieren wir die evolutionären Einflüsse auf die Kontaktoberflächen, um sie dadurch von den reinen Kristallkontakten zu unterscheiden.

Wir beschreiben die Methode im Detail und demonstrieren, dass sie effektiv zwischen biologischen und Kristallkontakten unterscheidet. Damit leistet die Methode einen wichtigen Beitrag zur Interpretation von Proteinkristallstrukturen und ermöglicht Rückschlüsse auf die biologische Anordnung von Proteinen in ihrer natürlichen Zellumgebung.

Die Methode wurde zunächst für lösliche Proteine entwickelt. In einem zweiten Teil zeigen wir, wie sie sich auch auf membrangebundene Proteine anwenden lässt. Dazu war zunächst die Zusammenstellung eines verifizierten Datensatzes von transmembranen Protein-Protein-Kontakten notwendig. Dies ist zugleich der erste umfangreiche und öffentliche Datensatz von verifizierten Transmembrankontakten. Anhand dieser Daten zeigen wir Prinzipien der Komplexbildung in der Membranumgebung und wie sich diese von derer in löslicher Umgebung unterscheiden.

Unsere Methode, genannt EPPIC, ist damit sowohl für lösliche als auch für Membranproteine anwendbar. Schließlich zeigen wir eine robuste Software-Implementierung der Methode sowohl als eigenständiges Programm als auch als web-basierter Online-Service, der somit der weltweiten Forschungsgemeinde zur Nutzung zur Verfügung steht.

1 Introduction

1.1 Background and Outline

Proteins are, together with nucleic acids, the most fundamental molecules of life. These molecular machines perform an enormous range of diverse functions that make life viable. Their incredible power to adapt to different functions is in a good part possible thanks to their ability to fold into very precise 3-dimensional shapes that can produce the most subtle variations and specialization to functions from conferring structural stability to tissues or catalyzing biochemical reactions to acting as logical gates at cell membranes.

Structural biology has thus become one of the most central branches of the biological sciences. Ever since Kendrew [1] solved the first protein structure in 1958 formidable technological advances have made possible an ever increasing accumulation of knowledge about the 3-dimensional structure of proteins. Crystallography has surely been at the center of this revolution, but other very important techniques like Nuclear Magnetic Resonance and Electron Microscopy have enabled a very wide range of studies around biological macromolecules. Indeed these two techniques, much younger than crystallography, offer still an enormous development potential. New emerging developments like solid state NMR or advances in EM detectors will surely open many more avenues in the structural biology research.

Crystallography has in any case made some of the greatest advancements in protein structural knowledge thanks to the ability to produce incredibly precise, atomic-

level detailed structures. The advent of synchrotron radiation sources together with advances in biotechnological techniques for protein preparation and crystallization and the immense increase in computational power have greatly pushed the limits of the complexity and number of macromolecules solved. Structural genomics initiatives have thus been able to make the high-throughput solution of structures a reality, something that was unimaginable only some years ago. Further automation is expected to produce even higher throughput pipelines that will keep increasing the rate of structures solved. Today the main repository of macromolecular structures, the Protein Data Bank (PDB) [2] accumulates more than 90,000 biological structures, a figure that has increased in the recent years by nearly 10,000 entries a year.

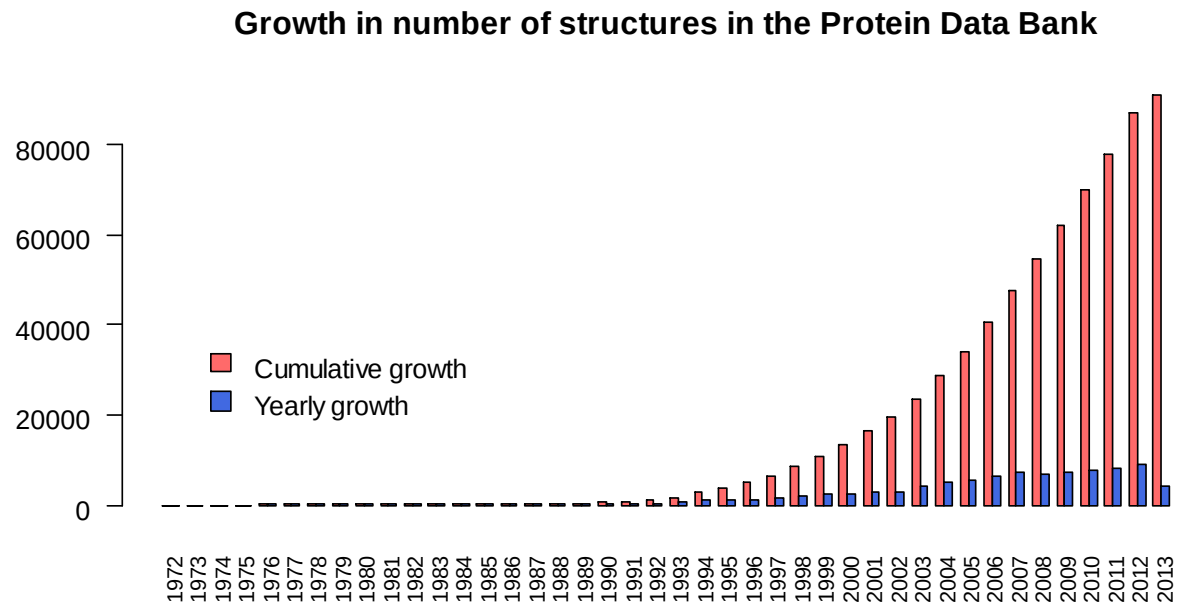


Figure 1 The growth in the number of structures in the Protein Data Bank since its creation

Free Electron Lasers (FELs), the next technological leap, are already under development and have produced the first promising results [3, 4]. These new X-ray radiation sources, producing ultra-short pulses many orders of magnitude more brilliant than synchrotron sources, are enabling possibilities such as serial nanocrystallography or single particle diffraction. Importantly this approach has the potential to eliminate radiation damage issues that affect synchrotron sources. Thus the new FEL revolution is expected to greatly contribute to the opening of new frontiers in structural biology.

With this wealth of data, the analysis of macromolecular structures has become ever more important. Structural bioinformatics thus has emerged as a very powerful

discipline that hopes to shed some light on the enormously complex world of protein structures. Great advances in the understanding and rationalization of structures have already been made, like the understanding that there exists only a limited set of folds in nature or the establishment of folding pathways that are ruled by the hydrophobic core collapse. Structural bioinformatics has also been very successful at solving many more practical problems like secondary structure prediction or even tertiary structure prediction by homology modeling.

Together with the structural data, another technological revolution that has contributed to develop the bioinformatics field is that of DNA sequencing. Since the completion of the human genome in 2001 [5, 6], the increase in throughput and lowering costs of DNA sequencing have happened at an even greater pace than that of computing technologies. Today it has become possible to sequence whole genomes in a few days for a fraction of the cost of some years ago. This has produced an explosion of sequence data that in turn fueled the appearance of fields such as comparative genomics, metagenomics and personalized medicine.

This thesis revolves around bringing together those two data explosions: structure and sequence. Much is to be gained from exploiting the sequence knowledge in order to shed light into the structural world. We try to bridge this gap and develop some new ideas utilizing the wealth of available data.

In particular we focus on a fundamental crystallographic problem: for all of its potential, one of the critical downsides of crystallography is that it does not provide any information about the specificity of the contacts present in the crystal lattice. As a protein goes from solution to crystalline phase, the quaternary structure arrangement is lost in the lattice. Being able to find out the original quaternary structure is nevertheless essential to understand proteins and their function.

We thus develop this thesis into the following chapters:

- **Introduction**: introducing the problem we are dealing with and its importance, crystal and biological contacts in crystal structures
- **Protein interface classification by evolutionary analysis**: the method developed to analyse crystal interfaces (published as Duarte et al 2012)
- **An analysis of interfaces in membrane protein structures**: the application of the method to interfaces of membrane proteins (Duarte et al, manuscript in preparation)
- **Conclusions**: including the application of the method to some interesting biological examples
- **Appendix**: two structural bioinformatics publications developed with the OWL framework that was essential for the realization of the project

1.2 A brief summary on protein quaternary structure

Polypeptide chains fold into well-defined 3-dimensional structures, in principle encoded in their amino acid sequence [7]. The fold of each of this single polypeptide chains is known as tertiary structure. However in order to perform their functions protein chains sometimes associate through non-covalent links to form oligomeric structures or to form complexes. This is referred to as the quaternary structure.

The formation of the oligomeric structures is mediated by protein-protein interfaces. Depending on their composition the oligomeric structures can be homo-oligomers if all subunits are of the same type or hetero-oligomers if different subunits associate together.

Following Monod [8] two kinds of interfaces or "two models of association" can occur in proteins:

- **Isologous** interfaces: the two partners bind in such a way that the two bonding surfaces are identical in both sides. This is the case of homodimers occurring on a 2-fold axis. They form closed symmetries and can only further associate by using other regions of their surfaces.
- **Heterologous** interfaces: the two partners interact through two different bonding surfaces. This kind of interfaces leads in general to infinite associations of helical polymers except in the case where a closed structure can be achieved where the two different sides of the interface are satisfied by "closing" the structure. This leads to homomers with cyclic point group symmetries Cn (n>=3).

Thus as Monod reasoned, homomers are necessarily symmetric: any heterologous interface not forming a closed symmetry would lead to infinite fiber assemblies.

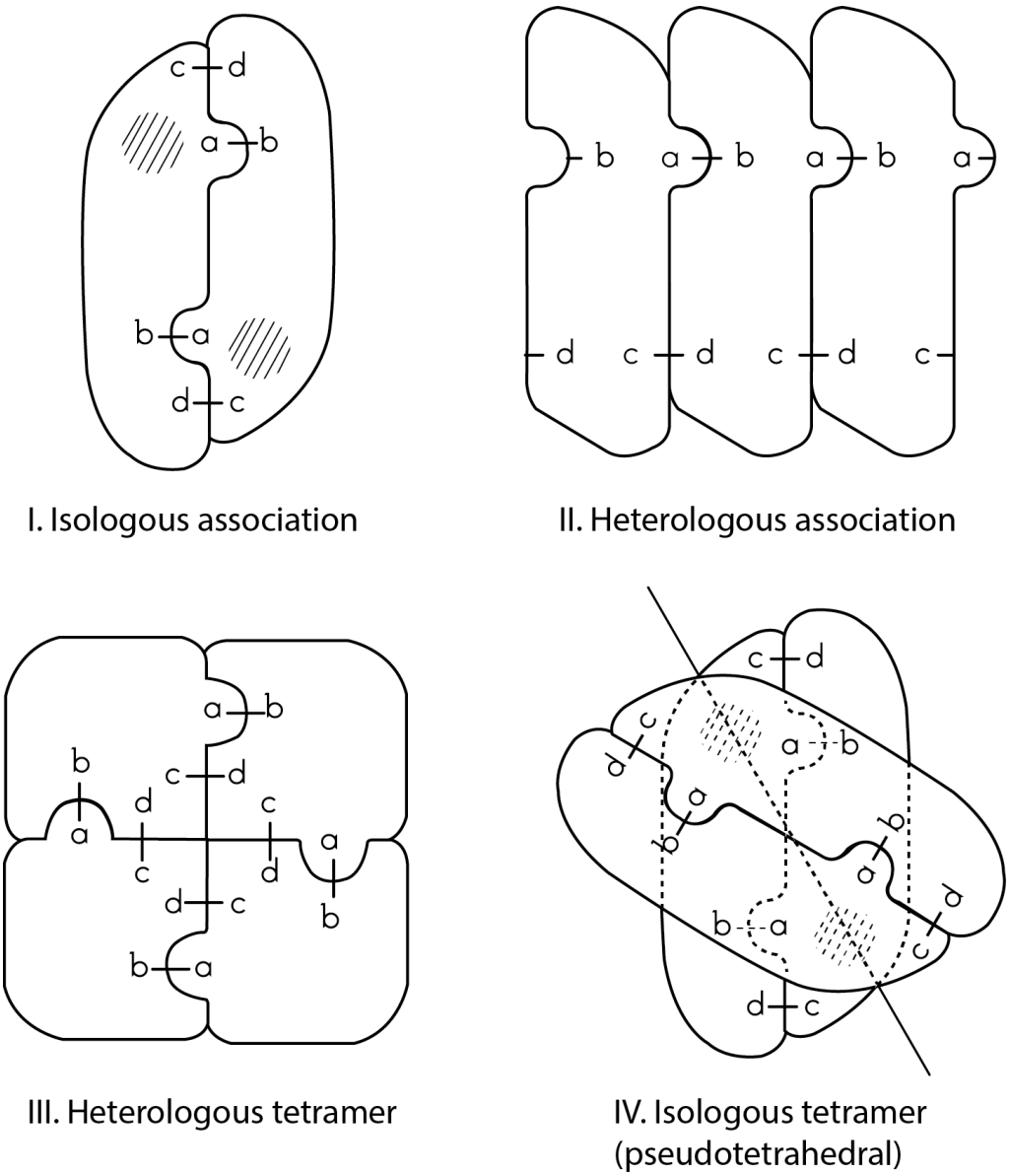


Figure 2 Different types of associations in proteins following Monod. Adapted from Monod et al [8]

Crystallography as already mentioned above is not able to disentangle the different interfaces present in the crystal lattice. It holds little information on what the correct biological assembly is. Thus a multitude of independent experimental techniques have been traditionally used to establish the oligomeric state of proteins. Usually they aim at establishing the molecular weight of the native protein and comparing it to the weight of the protomers known from sequence. Examples of such techniques are Size Exclusion Chromatography (gel filtration), Analytical Ultra Centrifugation or Light Scattering. Information about the oligomerization state can also be obtained from techniques such as NMR, small angle X-ray scattering (SAXS), chemical cross-linking and by mutagenesis studies in combination with other biophysical methods. Other new techniques are also emerging like Cross-linking Mass Spectrometry [9, 10] or oxidative footprinting [11].

1.3 Crystal lattice contacts and biological interfaces: a history

In the first decades of protein crystallography the issue of distinguishing crystal contacts from biological ones did not appear as prominent. In all cases the solution of a protein structure was preceded by years of biochemical characterization, so the quaternary structure was normally accurately known. Once the structure was solved finding out the contacts was in most cases a matter of visual inspection and relatively straight forward.

As years passed and more complex structures started to appear the problem became apparent. In 1995 Janin and Rodier were the first to openly describe the issue [12] and later published a further note on it [13]. By that time the PDB had already accumulated a not insignificant figure of more than 3000 structures. He could then compile statistics on the general geometrical features of crystal contacts in terms of their buried surface area. He was thus the first to note that crystal contacts tend to be small, while biological interfaces are usually much larger. The same observation was also confirmed by Carugo and Argos in their 1997 paper [14]. Janin's most relevant result was that by plotting the interface areas of interfaces found in crystals of monomeric proteins, an inverse exponential relationship could be seen in the area distribution. This offered the first method to a priori distinguish crystal from biological contacts, or specific versus non-specific ones in their nomenclature. By using their buried surface areas one could calculate a probability, based on the known statistics, of an interface to be specific or not.

Later the group of Janet Thornton was the first to provide an automated method to classify crystal and biological contacts. In 1998 Henrick and Thornton [15] presented the PQS service to automatically predict biological units from crystal structures. A mix of interface area, buried residues, solvation energy and number of salt and disulfide bridges were used to distinguish the two types of contacts. Later Ponstingl et al [16] used a knowledge-based statistical potential in order to try to differentiate the two types of contacts. By compiling statistics of pairwise atom contacts in known homodimer interfaces they could score a given interface and set a cut-off for classification. The method achieved an error rate of 12%, which was slightly better than the classification based purely on areas, achieving a 15% error rate on the same set of structures.

In 2001 in a pioneering study the same group performed an exploratory analysis of amino acid conservation in homodimers [17], which was hinted at as a possible way of solving the crystal versus biological interface conundrum. This was the first time that interfaces were looked at through the help of evolutionary data based on sequence analyses. In a very thorough analysis they measured the differential conservation signal between interface and other regions of the protein. As

conservation metrics they used a value termed *Cons*, a weighted sum of pairwise similarities among residues in an alignment column. They used it to compare the average conservation of the interface and other surface patches. They explored many different combinations: a) several definitions of interface residues: central residues and all residues, b) masked and unmasked ligand sites, c) sampling of surface patches with two different strategies: walking and picking.

Following their preliminary study, also in 2001 the same authors published a method for contact classification [18]. The study came out almost simultaneously with one by Elcock and McCammon [19] who aimed at solving the problem with similar conservation-based methods. In the case of Valdar and Thornton, the method applied was similar to their earlier one, except that they also combined their conservation measures with area in order to boost classification power. Elcock and McCammon introduced the usage of sequence entropy as a conservation metrics for this particular problem. However they used a simpler differential measure than Valdar's, by comparing the average entropy of whole interface versus the average of the rest of the protein's surface as a simple ratio.

Other studies followed shortly after from the Janin group with more in-depth analyses of geometric and physico-chemical properties of the two types of contacts. Chakrabarti and Janin [20] analyzed biological protein-protein interfaces in terms of size, geometry and amino acid composition. They introduced the idea of dividing the interfaces into a core and a rim region, finding that the rim has a similar composition to the rest of the protein surface while the core composition differs from it. The core was defined as those residues that bury at least one of their atoms upon interface formation. Bahadur et al [21] followed this up by noting that crystal contact interfaces seem to have similar amino acid composition to that of the rest of the surface, while biological interfaces are more hydrophobic and have more fully buried atoms. Furthermore they came up with a score combining those two observations that was one of the most effective to that date at distinguishing specific from non-specific contacts. Importantly for the first time they recognized that a crystal interface classification method needs to be somehow decoupled from the buried area parameter. The total buried surface area although a blessing for classification as seen above can also constitute a hindrance by hiding the contribution of other parameters to the classifier. With this idea in mind they used a previously compiled dataset of crystal contacts filtering out those contacts that were burying less than 400 $Å^2$ per protomer.

In 2005 following some of the ideas of the Chakrabarti and Bahadur papers, Guharoy and Chakrabarti [22] added conservation to the mix making the next contribution to the evolutionary methods, after those of Valdar et al and Elcock et al above. Using sequence entropies as evolutionary metrics they compared the average entropies of

core and rims as defined previously by Chakrabarti. The rationale behind it being that the comparison to surfaces might introduce bias from other conserved surface patches like ligand binding sites, thus they argued that a core versus rim approach can give better signal.

Not much later, in 2007, a very different method came into the game [23]. Krissinel et al created the PISA software package for completely automated prediction of biological units from crystal structures. Their ideas were along the lines of analysis of geometrical and physico-chemical properties of interfaces, but they were integrated in a formal frame of interface energetic stability calculations from first principles. Starting from the full thermodynamic equations for the dissociation equilibrium of protein complexes, the authors go through a series of simplifications that finally lead to two main contributions: a) the enthalpic one containing contributions from ΔASAs, counts of hydrogen bonds, salt and disulfide bridges; b) the entropic one containing contributions from mass, moments of inertia, symmetry numbers and again ΔASA. The model makes quite a lot of assumptions and needs a few unknown empirical parameters, which are found by optimizing against their training set of bona fide biological and crystal contacts.

Despite the coarse modelling the PISA method could achieve a very good performance. Key to its success was the assembly algorithm that could enumerate all possible assemblies present in the crystal producing a list of plausible biological units. The algorithm favors high symmetry and larger oligomeric ensembles and eliminates parallel interfaces, i.e. those that would lead to infinite assemblies. Last but not least they provided a working implementation including an easy-to-use web server that contributed enormously to the popularity of the project, so much that it has gradually become the *de facto* standard in the field.

A few other approaches have also been tried. Nearly all of them revolve around the ideas presented above and many times try to group the different indicators together, being them geometrical, physico-chemical properties or evolutionary ones. The indicators are then fed into machine learning algorithms in order to automatically classify the interfaces by training with known sets of biological/crystal contacts. The first of such methods to appear (2006) was that of Zhu et al [24] where a Support Vector Machine (SVM) was trained with 6 different interface parameters combining areas, amino acid properties, conservation and geometry of interface. The study of Bernauer et al [25] also came up with one such method where as many as 84 geometrical and amino acid properties-related parameters (reduced to 27 in the most optimal set) are fed into a SVM. A third method was that of Mitra et al [26] where a Bayes classifier was used instead of a SVM, this time with a combination of geometrical and physico-chemical parameters. In this case the authors go beyond

pairwise interface classification by looking at the point group symmetry and coming up with full biological unit predictions.

A very important issue affecting the machine learning methods is the predominance of the interface area as the most effective classifying indicator. Together with that, highly interdependent indicators many times correlated with interface area, were often used in the machine learning procedures. Another important issue is that of over-fitting, especially taking into account the scarcity of data available for training.

An interesting analysis described in 2008 by Xu et al [27] looked at the problem from a totally different perspective than any of those above. They realized that the wealth of structures solved in the PDB could be exploited to look at conserved crystal interfaces of homologous proteins in different crystal forms, under the assumption that such conserved interfaces must be biological. The method can be a powerful validation tool in some cases or even a discovery tool for potential undetected biological interfaces. Of course its power is limited by the availability of redundant crystal forms of the same structure in the PDB.

In summary many aspects of the crystal interface classification problem have already been explored and have greatly deepened our understanding of the issue. Some important conclusions can be extracted:

- Crystal interfaces tend to be small and biological ones tend to be large. Thus for the extreme cases, classification is more or less trivial, but a difficult region of areas exist where disentangling the two becomes a lot more challenging.
- Geometrical and packing measures can be used quite effectively to classify the interfaces.
- Building assemblies and looking at their size and symmetry provides a lot of additional valuable information about the interfaces.
- Evolutionary data coming from sequences can provide very good hints on the biologicity of the interfaces.

Our focus in this thesis is that of classification through evolutionary data as we feel that from all aspects studied so far it is the one with the most opportunities for development, but still not exploited to its full potential. We review thus in the next section some of the different ways of "measuring" evolution in particular in their relation to protein structure.

1.4 Evolution and selection pressure metrics

A multiple sequence alignment provides a picture of the results of millions of year of evolution in a single snapshot. When in combination with protein structures the

kind of information contained in the multiple sequence alignment can greatly enhance the value of the structural data, helping one to understand different aspects of the protein fold and function.

The aligned positions of the MSA correspond to precise locations in the 3-dimensional structure of a particular protein family. Thus the variability observed for one of such columns carries information regarding the allowed variability of amino acids in the particular 3-dimensional location. Amino acid substitutions that lead to a disruption of the fold or lose of binding to a particular cofactor or ligand will thus not be present in a MSA since the substitution would disrupt the protein's function.

Quantification of the variability of the alignment columns is thus essential if one wants to measure these effects in practice. This problem has been subject of investigation for already quite a few decades. A comprehensive review of all these developments falls out of the scope of this thesis, however we will briefly introduce some of the quantitative methods that have been devised to study the problem in the context of structural applications.

We will divide this section into two parts corresponding to the two subdivisions within the problem: selection of sequence homologs for the MSA and calculation of a selection pressure metric.

1.4.1 Selection of sequence homologs

Most strategies for selection of sequence homologs have usually aimed at maximizing the amount of sequences, trying to gather as many related sequence as possible even if only remote sequence homology exists. The rationale behind it being that protein structures are known to be well conserved even at low sequence identities. The limited availability of sequence data, especially before the advent of next generation sequencing, also justified in many occasions this kind of strategy.

The HSSP database [28] by Sander and Schneider was one of the earliest attempts at comprehensively associating sequence evolution information to known protein structures in the PDB. At the time of the study (1991) a limited amount of data was available: around 500 protein structures and 12000 sequences were known. The study tries to bridge the gap between the two worlds in order to infer more structural data from homology. Sander and Schneider reasoned that the threshold of sequence similarity for structural homology depends on the alignment length. By using the available data they provide an estimate for the threshold of sequence identity that implies structure homology for different sequence lengths, finding that at a length of 10 residues structural homology can only be seen at 80% sequence identity, whilst for sequences longer than 80 residues structural homology is already

mandated by a 25% sequence identity level. Based on these estimates they provide a database of alignments for all available protein structures using a variable cut-off depending on their sequence length. Nearly all protein structures these days exceed the 80 residues length limit and thus the generous 25% identity limit is used for many proteins in the current HSSP database.

Later other methods came along that put a lot of effort into the evolution metrics but did not try to justify so rigorously the homolog search strategy and mostly aimed at gathering as many sequences as possible to infer more information from the broad alignments. For instance the MP-consurf method [29] used the Smith-Waterman algorithm to collect sequences from the SwissProt database with an E-value cut-off of 0.05 which results in a very broad sequence spectrum. Later the same group developed an improved version of the method [30], named ConSurf-HSSP, where the selection of homologs was left to the HSSP database. Valdar and Thornton [18] used PSI-blast with a maximum of 20 iterations and an E-value cut-off for inclusion of E-40 resulting in sequences with very low identities, as low as 5%, to the queried one.

Other variations of these search strategies have also appeared. The search for remote homologs has always been a central issue. Earlier methods like BLAST [31] or implementation of exact algorithms like Needleman-Wunsch [32], were based purely on pairwise relationships and used substitution matrix models [33] to find related sequences. Later more sophisticated context-specific methods appeared, some of them based on sequence profiles like PSI-BLAST [34] and others based on Hidden Markov Models like HMMER [35]. These methods gained in sensitivity by gathering context information from a seed alignment, i.e. specific evolution patterns for a particular protein family, and were able to find much more distant homologs. Development has not stopped there and improved versions of those methods have recently been published, like DELTA-BLAST [36], HMMER3 [37] or HHblits [38].

1.4.2 Measuring evolution

Measuring residue conservation from a Multiple Sequence Alignment has been a matter of study for a long time, for instance a very comprehensive review was written by Valdar in 2002 [39]. Briefly, the problem can be further subdivided into separate issues:

- Weighting of sequences: the alignments can contain sequences that are very similar to each other and can thus bias the calculations. Two possible approaches to compensate for this redundancy are pre-filtering by sequence identity clustering or weighting of sequences by similarity.
- Variability measure: based on amino acid variance measures, on Shannon's entropy (with different groupings of amino acids) or on substitution matrices.

- Normalization: one can decide whether to use the scores as calculated or to normalize them in order to make them more homogeneous. One possible approach is that of normalizing by the average conservation from the whole alignment.
- Treatment of gaps: a very important issue is how to handle the gaps in the conservation measures. Clearly gaps are not just another symbol that can be handled as other amino acids, but require a special treatment. Ignoring them altogether also leads to biasing the conservation scores. Some possible approaches here are: weighting down columns by gappyness or ignoring columns with a gap content exceeding a certain threshold.

Due to the complexity of the issue and the many different parameters involved, groups have usually centered on a particular measure. Studies have also been performed that compared many of these different measures like that of Pei and Grishin [40]. The authors not only compared the different variability measures but also different ways of aligning the sequences: curated SMART alignments [41], ClustalW sequence alignments [42] or FSSP structural alignments [43]. Unsurprisingly high correlation could be observed between the different measures. Pei and Grishin could also use the conservation measures as quality metrics for the alignments and found out that not much difference could be seen between the SMART and ClustalW alignments, while the structure-based FSSP ones were somewhat more accurate.

Phylogenetic measures can also be introduced that somehow model the evolutionary history of the sequences. This was central to some studies like ConSurf [30] or the Evolutionary Trace method [44]. There the phylogeny of the sequences is used as a guide to model sequence evolution, leading to in-principle more accurate metrics for selection pressure. Two potential weaknesses of these methods, however, are their need for many assumptions for modeling sequence evolution and the fact that phylogenetic trees are calculated, in the lack of other data, also from sequence similarities.

A further type of measure, not discussed above, is based on nucleotide sequences instead of protein amino acid sequences. These methods are popular in the Molecular Evolution field. The introduction of nucleotide sequences opens the door to more evolutionary information from the codons that encode for amino acids. Since the genetic code is degenerate, additional variability not seen in the amino acid sequence can be detected.

Very generally the different coding sequence-based models try to correct for different evolutionary rates occurring in different protein families. One possible such measure is the Ka/Ks ratio [45]: Ka represents the rate of non-synonymous

substitutions per non-synonymous site, while Ks is the rate of synonymous substitutions per synonymous site. The reasoning behind it is that the synonymous mutation rate gives an idea of the background drift of neutral evolution, since the silent mutations not involving amino acid substitutions do not affect the fitness of the particular protein encoded by the DNA sequence. The ratio of Ka over Ks thus provides a measure of the selection pressure acting on a sequence with a correction for the neutral drift. Thus the metric is in principle able to better convey the evolutionary patterns of the sequences. However the necessary and complicated models for multiple substitutions require themselves many assumptions and add to the complexity of the method, compared to the more naïve –but clear– earlier methods purely based on protein sequences.

1.5 Initial motivation and previous explorations

The project that spawned this thesis was started by my supervisor Guido Capitani and my former colleague Martin Schärer who completed his doctoral thesis in the "Crystallography and Structural Bioinformatics" Capitani team at the Paul Scherrer Institute.

One of the crystallographic topics of the team was the structural characterization of the Type 1 Pilus from the uropathogenic Escherichia Coli bacterium, in collaboration with the Glockshuber lab at ETH Zurich. The Pilus assembly machinery is composed of many different proteins that assemble at the outer membrane of these bacteria. One of the complexes that are involved in the pilus assembly is that of the FimD-FimC-FimH, described by Nishiyama et al [46] in the frame of a collaboration involving as well the Wüthrich lab at ETHZ. In the course of that study an ambiguity was found in the assembly of the ternary complex: two different choices of asymmetric unit, crystallographically equivalent, led to biologically different assemblies of the complex. The ambiguity had to be resolved at the time by site-directed mutagenesis and biophysical measurements.

Inspired by that problem the team initiated the CRK project, which lead eventually to the publication of the Schärer et al paper in 2010 [47]. CRK stands for Core-Rim Ka/Ks ratio and focuses on solving the problem of distinguishing biologically-relevant interfaces from crystal contacts by using an evolutionary approach. The study introduced a new approach by borrowing concepts from the Molecular Evolution field, which had not been used in this context. The Ka/Ks ratio was used as selection pressure metric, which as presented above, measures evolutionary rates by using DNA coding sequences. By introducing this novel methodology and combining it with an improved definition of interface core residues, the method was very successful at classifying the interfaces, achieving an 84% accuracy overall.

Two key factors contributed to the success of that project:

- The definition of core residue: interface core residues were defined as those burying 95% of the ASA upon complexation, a definition that was a lot more stringent than those employed in previous studies measuring differential selection pressure of core and rim of interfaces.
- The careful selection and filtering of sequences in the alignments: strict criteria were used where closed sequence homologs of the protein structures were chosen in order to be able to measure properly the Ka/Ks ratios. This measure can suffer from issues like Ks saturation if too distant sequences are grouped together in the alignment.

Surely the abundance of sequence data as compared to the earlier work also contributed to boost the performance of the method. Finally, a careful treatment of the crystallographic data and checking of its quality and correctness contributed greatly to produce good results.

All in all the study demonstrated the viability of the method and served as a starting point for the development of the ideas that came in the Duarte et al [48] paper presented in the next section. The method was implemented in a few scripts and provided a proof-of-concept but was not distribution-ready. Thus one of the objectives deriving from it was that of a robust distributable implementation together with the development of a Graphical User Interface in the form of a web server.

1.6 A final note on software

Since the start of the present thesis and originating from my working period at the Max Planck Institute for Molecular Genetics in Berlin, one of the central focuses of the scientific work was the development of high quality and sustainable software.

In any scientific endeavor careful experimentation and proper planning and structuring are essential to the success of a project. This applies equally well to scientific projects that are based on software development. Often the development of scientific software occurs in a less than optimal manner constrained by time, lack of resources and above all by the rapid pace of development of the new scientific ideas in unexplored territories. In this kind of context proper software development practices are not always respected. It is our belief that software engineering practices should always be applied and that only this kind of "careful experimentation" can lead to correct analyses and new discoveries.

The projects presented in this thesis were made possible by the previous development of the Otto Warburg Java library (OWL) for structural bioinformatics

(http://www.bioinformatics.org/owl), created by myself and colleagues in Michael Lappe's group at the Max Planck Institute of Molecular Genetics in Berlin. The group focused on protein structure studies, especially structure representation as residue interaction graphs or networks. The OWL software library was developed in order to enable the different analyses performed in the group. Some of its features are: input and output of PDB structures, implementation of a fast residue interaction graph decomposition algorithm, contact map handling or implementation of a homology modeling pipeline.

Several publications were facilitated by the OWL library [49–54]. Additionally it enabled the participation of the Lappe group in the Critical Assessment of Structure Prediction (CASP) experiment in its 8th and 9th edition.

During my thesis work the library was greatly expanded and made much more versatile. For instance I implemented fast algorithms for the reconstruction of the crystal lattice from the asymmetric unit, fast inter-chain contacts calculation or a pipeline for sequence homolog searching including calculation of conservation scores.

In the Appendix of this thesis two of the publications where I was involved and that employed the OWL library are presented. They both were published during my working period at the Paul Scherrer Institute.

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC: **A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis**. *Nature* 1958, **181**:662–666.

2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic acids research* 2000, **28**:235–42.

3. Chapman HN, Fromme P, Barty A, White TA, Kirian RA, Aquila A, Hunter MS, Schulz J, DePonte DP, Weierstall U, Doak RB, Maia FRNC, Martin A V, Schlichting I, Lomb L, Coppola N, Shoeman RL, Epp SW, Hartmann R, Rolles D, Rudenko A, Foucar L, Kimmel N, Weidenspointner G, Holl P, Liang M, Barthelmess M, Caleman C, Boutet S, Bogan MJ, Krzywinski J, Bostedt C, Bajt S, Gumprecht L, Rudek B, Erk B, Schmidt C, Hömke A, Reich C, Pietschner D, Strüder L, Hauser G, Gorke H, Ullrich J, Herrmann S, Schaller G, Schopper F, Soltau H, Kühnel K-U, Messerschmidt M, Bozek JD, Hau-Riege SP, Frank M, Hampton CY, Sierra RG, Starodub D, Williams GJ, Hajdu J, Timneanu N, Seibert MM, Andreasson J, Rocker A, Jönsson O, Svenda M, Stern S, Nass K, Andritschke R, Schröter C-D, Krasniqi F, Bott M, Schmidt KE, Wang X, Grotjohann I, Holton JM, Barends TRM, Neutze R, Marchesini S, Fromme

R, Schorb S, Rupp D, Adolph M, Gorkhover T, Andersson I, Hirsemann H, Potdevin G, Graafsma H, Nilsson B, Spence JCH: **Femtosecond X-ray protein nanocrystallography.** *Nature* 2011, **470**:73–7.

4. Seibert MM, Ekeberg T, Maia FRNC, Svenda M, Andreasson J, Jönsson O, Odić D, Iwan B, Rocker A, Westphal D, Hantke M, DePonte DP, Barty A, Schulz J, Gumprecht L, Coppola N, Aquila A, Liang M, White TA, Martin A, Caleman C, Stern S, Abergel C, Seltzer V, Claverie J-M, Bostedt C, Bozek JD, Boutet S, Miahnahri AA, Messerschmidt M, Krzywinski J, Williams G, Hodgson KO, Bogan MJ, Hampton CY, Sierra RG, Starodub D, Andersson I, Bajt S, Barthelmess M, Spence JCH, Fromme P, Weierstall U, Kirian R, Hunter M, Doak RB, Marchesini S, Hau-Riege SP, Frank M, Shoeman RL, Lomb L, Epp SW, Hartmann R, Rolles D, Rudenko A, Schmidt C, Foucar L, Kimmel N, Holl P, Rudek B, Erk B, Hömke A, Reich C, Pietschner D, Weidenspointner G, Strüder L, Hauser G, Gorke H, Ullrich J, Schlichting I, Herrmann S, Schaller G, Schopper F, Soltau H, Kühnel K-U, Andritschke R, Schröter C-D, Krasniqi F, Bott M, Schorb S, Rupp D, Adolph M, Gorkhover T, Hirsemann H, Potdevin G, Graafsma H, Nilsson B, Chapman HN, Hajdu J: **Single mimivirus particles intercepted and imaged with an X-ray laser.** *Nature* 2011, **470**:78–81.

5. Venter JC, Adams MD, Myers EW, et al.: **The sequence of the human genome.** *Science (New York, N.Y.)* 2001, **291**:1304–51.

6. Lander ES, Linton LM, Birren B, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

7. Anfinsen CB, Haber E, Sela M, White FH: **The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.** *Proceedings of the National Academy of Sciences of the United States of America* 1961, **47**:1309–14.

8. Monod J, Wyman J, Changeux JP: **On the nature of allosteric transitions: a plausible model.** *Journal of molecular biology* 1965, **12**:88–118.

9. Zhang H, Tang X, Munske GR, Tolic N, Anderson GA, Bruce JE: **Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry.** *Molecular & cellular proteomics : MCP* 2009, **8**:409–20.

10. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R: **Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics.** *Molecular & cellular proteomics : MCP* 2010, **9**:1634–49.

11. Guan J-Q, Chance MR: **Structural proteomics of macromolecular assemblies using oxidative footprinting and mass spectrometry.** *Trends in biochemical sciences* 2005, **30**:583–92.

12. Janin J, Rodier F: **Protein-protein interaction at crystal contacts.** *Proteins* 1995, **23**:580–7.

13. Janin J: **Specific versus non-specific contacts in protein crystals.** *Nature structural biology* 1997, **4**:973–4.

14. Carugo O, Argos P: **Protein-protein crystal-packing contacts.** *Protein science : a publication of the Protein Society* 1997, **6**:2261–3.

15. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server.** *Trends in biochemical sciences* 1998, **23**:358–61.

16. Ponstingl H, Henrick K, Thornton JM: **Discriminating between homodimeric and monomeric proteins in the crystalline state.** *Proteins* 2000, **41**:47–57.

17. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42**:108–124.

18. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *Journal of molecular biology* 2001, **313**:399–416.

19. Elcock AH, McCammon JA: **Identification of protein oligomerization states by analysis of interface conservation.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:2990–4.

20. Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites.** *Proteins* 2002, **47**:334–43.

21. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces.** *Journal of molecular biology* 2004, **336**:943–55.

22. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:15447–52.

23. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state.** *Journal of Molecular Biology* 2007, **372**:774–797.

24. Zhu H, Domingues FS, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types.** *BMC bioinformatics* 2006, **7**:27.

25. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A: **DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.** *Bioinformatics (Oxford, England)* 2008, **24**:652–8.

26. Mitra P, Pal D: **Combining Bayes Classification and Point Group Symmetry under Boolean Framework for Enhanced Protein Quaternary Structure Inference.** *Structure London England 1993* 2011, **19**:304–12.

27. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL: **Statistical analysis of interface similarity in crystals of homologous proteins.** *Journal of molecular biology* 2008, **381**:487–507.

28. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**:56–68.

29. Armon A, Graur D, Ben-Tal N: **ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information.** *Journal of molecular biology* 2001, **307**:447–63.

30. Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N: **The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures.** *Proteins* 2005, **58**:610–617.

31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215**:403–10.

32. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *Journal of Molecular Biology* 1970, **48**:443–453.

33. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins**. *Atlas of protein sequence and structure* 1978, **5**:345–352.

34. Altschul S: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**:3389–3402.

35. Eddy SR: **Profile hidden Markov models.** *Bioinformatics (Oxford, England)* 1998, **14**:755–63.

36. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL: **Domain enhanced lookup time accelerated BLAST.** *Biology direct* 2012, **7**:12.

37. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS computational biology* 2011, **7**:e1002195.

38. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nature methods* 2012, **9**:173–5.

39. Valdar WSJ: **Scoring residue conservation.** *Proteins* 2002, **48**:227–41.

40. Pei J, Grishin N V: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics (Oxford, England)* 2001, **17**:700–12.

41. Schultz J, Milpetz F, Bork P, Ponting CP: **Colloquium Paper: SMART, a simple modular architecture research tool: Identification of signaling domains**. *Proceedings of the National Academy of Sciences* 1998, **95**:5857–5864.

42. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic acids research* 1994, **22**:4673–80.

43. Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins.** *Nucleic acids research* 1996, **24**:206–9.

44. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *Journal of molecular biology* 1996, **257**:342–58.

45. Li WH, Wu CI, Luo CC: **A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.** *Molecular biology and evolution* 1985, **2**:150–74.

46. Nishiyama M, Horst R, Eidam O, Herrmann T, Ignatov O, Vetsch M, Bettendorff P, Jelesarov I, Grütter MG, Wüthrich K, Glockshuber R, Capitani G: **Structural basis of chaperone-subunit complex recognition by the type 1 pilus assembly platform FimD.** *The EMBO journal* 2005, **24**:2075–86.

47. Schärer MA, Grütter MG, Capitani G: **CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts.** *Proteins* 2010, **78**:2707–2713.

48. Duarte JM, Srebniak A, Schärer M a, Capitani G: **Protein interface classification by evolutionary analysis.** *BMC bioinformatics* 2012, **13**:334.

49. Bolser DM, Filippis I, Stehr H, Duarte J, Lappe M: **Residue contact-count potentials are as effective as residue-residue contact-type potentials for ranking protein decoys.** *BMC structural biology* 2008, **8**:53.

50. Sathyapriya R, Duarte JM, Stehr H, Filippis I, Lappe M: **Defining an Essence of Structure Determining Residue Contacts in Proteins**. *PLoS Computational Biology* 2009, **5**:10.

51. Duarte JM, Sathyapriya R, Stehr H, Filippis I, Lappe M: **Optimal contact definition for reconstruction of contact maps.** *BMC bioinformatics* 2010, **11**:283.

52. Stehr H, Duarte JM, Lappe M, Bhak J, Bolser DM: **PDBWiki: added value through community annotation of the Protein Data Bank**. *Database* 2010, **2010**:baq009–baq009.

53. Vehlow C, Stehr H, Winkelmann M, Duarte JM, Petzold L, Dinse J, Lappe M: **CMView: Interactive contact map visualization and analysis.** *Bioinformatics (Oxford, England)* 2011:2–3.

54. Stehr H, Jang S-HJ, Duarte JM, Wierling C, Lehrach H, Lappe M, Lange BM: **The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors.** *Molecular Cancer* 2011, **10**:54.

2 Protein interface classification by evolutionary analysis

Duarte et al. BMC Bioinformatics 2012, 13:334
http://www.biomedcentral.com/1471-2105/13/334

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Protein interface classification by evolutionary analysis

Jose M Duarte[1], Adam Srebniak[2], Martin A Schärer[1,3] and Guido Capitani[1*]

Abstract

Background: Distinguishing biologically relevant interfaces from lattice contacts in protein crystals is a fundamental problem in structural biology. Despite efforts towards the computational prediction of interface character, many issues are still unresolved.

Results: We present here a protein-protein interface classifier that relies on evolutionary data to detect the biological character of interfaces. The classifier uses a simple geometric measure, number of core residues, and two evolutionary indicators based on the sequence entropy of homolog sequences. Both aim at detecting differential selection pressure between interface core and rim or rest of surface. The core residues, defined as fully buried residues (>95% burial), appear to be fundamental determinants of biological interfaces: their number is in itself a powerful discriminator of interface character and together with the evolutionary measures it is able to clearly distinguish evolved biological contacts from crystal ones. We demonstrate that this definition of core residues leads to distinctively better results than earlier definitions from the literature. The stringent selection and quality filtering of structural and sequence data was key to the success of the method. Most importantly we demonstrate that a more conservative selection of homolog sequences - with relatively high sequence identities to the query - is able to produce a clearer signal than previous attempts.

Conclusions: An evolutionary approach like the one presented here is key to the advancement of the field, which so far was missing an effective method exploiting the evolutionary character of protein interfaces. Its coverage and performance will only improve over time thanks to the incessant growth of sequence databases. Currently our method reaches an accuracy of 89% in classifying interfaces of the Ponstingl 2003 datasets and it lends itself to a variety of useful applications in structural biology and bioinformatics. We made the corresponding software implementation available to the community as an easy-to-use graphical web interface at http://www.eppic-web.org.

Keywords: Protein structure, Protein-protein interfaces, Crystal interfaces, Classification, Evolutionary, Core residues, Web server

Background

Protein crystal lattices contain two kinds of interfaces: biological ones (as present in physiological conditions) and crystal packing ones (non-specific), indistinguishable by crystallographic means. Traditionally they have been assigned by visual inspection alone, but their identification has increasingly become a challenge due to the sheer complexity of the macromolecular objects that modern structural biology tackles nowadays. A series of breakthroughs in protein production and structure determination techniques, especially in protein crystallography, nuclear magnetic resonance and electron microscopy, have enabled researchers to solve the structure of macromolecular complexes and oligomeric proteins of very large size, sometimes composed of many copies of different kinds of subunits. Prominent examples in this respect are for instance fatty acid synthase [1] and the recently solved immunoproteasome [2]. Another important trend is the increasing automation of the structure determination pipeline through structural genomics efforts, often producing protein structures before thorough biochemical characterization. Reliable computational tools are thus needed to decide which interfaces are the biologically relevant ones and consequently what is the biological assembly in the crystal. The need for such tools is not limited

* Correspondence: guido.capitani@psi.ch
[1]Paul Scherrer Institut, Villigen CH-5232, Switzerland
Full list of author information is available at the end of the article

to crystallography: integrated approaches merging electron microscopy, proteomics and crystallography are being employed to tackle very complex entities such as the nuclear pore complex [3,4]: there, researchers determine the structures of individual components in order to fit them into a lower resolution global electron density map derived from electron microscopy data. It is vital, in order to obtain a correct fit, to know if the assemblies of the components obtained by crystallography are biologically relevant.

In the last fifteen years several computational methods have been developed to distinguish biological interfaces from crystal contacts. The first of them relied on interface area analysis [5] and was followed by approaches based on sequence conservation [6-8], combination of geometrical and other properties such as conservation via machine learning [9-11] and thermodynamic estimation of interface stability [12]. This last method, implemented in the PISA server, proved to be the most successful and is the current *de facto* standard in the field. An interesting approach, PROTCID [13,14], infers information about the biological significance of interfaces from their presence in multiple crystal forms of the same protein (if available).

In this article we present an integrated approach to the problem that relies on evolutionary analysis of the interfaces and on a novel geometric criterion. In a previous, proof-of-concept work [15] we employed *Ka/Ks* ratios as the evolutionary metrics for the selection pressure acting on protein-protein interfaces in crystals. *Ka/Ks* ratios are a well-established tool in the field of molecular evolution: they measure the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site in a multiple alignment of coding sequences [16]. We compared the *Ka/Ks* ratio averages of interface rim and core sets to detect if the selection pressure acting on core residues was significantly stronger than that of rim residues. This approach had three main limitations: first, its recall was limited since in many cases not enough homologs could be found to run a significant *Ka/Ks* ratio estimation. Second, *Ka/Ks* value estimation was slow, bringing the duration of most runs up to several hours. Third, no easy-to-use public implementation was available.

Our new approach, named EPPIC (Evolutionary Protein-Protein Interface Classifier), overcomes all the above limitations, introduces two novel criteria for detecting biological contacts and, most importantly, achieves a very high level of accuracy. Additionally we implemented it in a robust freely available software package and offer it to the community in an easy-to-use graphical web interface.

Results and discussion

EPPIC is an approach for distinguishing biological interfaces from lattice contacts in crystal structures using evolutionary information from protein sequences. Some early attempts [7,8] in this direction, using sequence entropies as metrics for selection pressure, did not achieve levels of accuracy high enough to make them competitive with methods like PISA, which estimates the thermodynamic stability of an interface to predict whether it should exist in solution (biological interface) or only in the crystalline state (crystal contact). To date, PISA is the *de facto* standard to address the biological interface versus crystal contact issue and to predict the biologically relevant assembly of protein structures. Since PISA makes no use of sequence information, complementary methods that employ the wealth of sequence data available are particularly needed, especially as the size of biological sequence databases has increased exponentially in the last years and will only keep increasing further in the near future.

Our recent approach [15] aimed at demonstrating the feasibility of an evolution-based method measuring interface selection pressure at the coding-sequence level. Having achieved that goal, we set out to develop a completely new, more powerful and general approach to the problem, overcoming the limitations described in the introduction. First of all, we introduced a new geometric analysis criterion, based on the number of core residues in an interface, which represents by itself a powerful predictor of interface character. This allows us to formulate an interface assignment even when not enough homologs to the query are available for evolutionary analysis. Second, we have re-evaluated the use of sequence entropies instead of *Ka/Ks* ratios as a metrics for selection pressure. We found out that, with stringent criteria for homolog selection, better redundancy reduction of sequences and thanks to the increasing amount of sequences currently available, we could reach a better performance than that achieved with *Ka/Ks* ratios. The usage of entropies brings the advantage of making calculations much faster but also of simplifying the computational workflow. Third, we have introduced a new way to exploit the difference in selection pressure between interface and surface residues. Comparing the average sequence entropies of interface and non-interface residues is an approach pioneered by Elcock & McCammon [7]. That early attempt, however, was limited by the small size of sequence databases at the time and most importantly by biasing factors acting on surface residues, *e.g.* allosteric binding sites, unknown interfaces to other partners, external active sites and the like. We modified that approach substantially, first of all by comparing only interface core residues with surface residues, and by introducing a random pooling of surface residues

that enables us to compute more statistically robust scores of selection pressure acting on the interface core residues with respect to the surface "baseline". A similar surface sampling approach was also used successfully by Valdar and Thornton [17] in order to analyze conservation in a small set of homodimer interfaces.

The above three criteria, combined with further statistical considerations on interface area, allow us to achieve a performance of 89% accuracy on a minimally modified version of the Ponstingl 2003 dataset [18] as compared with 84% accuracy achieved by PISA on the same interfaces.

Compilation and annotation of new reference datasets

An important issue we tackled in this study was that of reference datasets of crystal contacts and biological interfaces. We identified this as one of the most important issues in the computational prediction of interface character and believe this particular problem has not received enough attention in previous studies. Experimental methods for oligomeric state determination are themselves prone to artifacts and it is rather common in the literature to find debated assignments, based on contradictory experimental data. It is thus essential that the data to be used for method developing and benchmarking have 100% clear experimental backing. The crystallographic accuracy of the structures is also vital: we realized that some of the most frequently used datasets in the literature contained some structures not following the most stringent crystallographic quality criteria, since they were solved many years ago and predated the use of quality measures such as the free R-factor [19].

Another important issue that has been mostly neglected is the distribution of areas of the interfaces used to train or benchmark classifier algorithms. As demonstrated already by Janin [5], an exponential decay relationship exists in the distribution of areas of crystal interfaces: the bulk of the crystal interfaces known to date have areas below 1000 $Å^2$ with very few representatives above that value. It is also well known that biological interfaces on the contrary tend to exhibit large areas [20], with a majority of cases from 1000 $Å^2$ and above. An overlap region exists where both kind of interfaces are frequent in the area values of approximately 800 $Å^2$ to 2000 $Å^2$. Thus an interface-classifying method should always take this into account and use this area distribution as a baseline for predictions. As Ponstingl [21] already noted, a simple classifier based on area alone achieved high accuracy in interface assignment. In introducing our own reference datasets we prioritized having a distribution of areas that is out of the trivially classifiable region. This issue was first recognized and partly addressed in the work of Bahadur *et al*. [22], where they included a crystal interface in their dataset only if its total buried area was above 400 $Å^2$.

We thus created our own reference datasets, adopting a three-fold strategy: 1) only use entries for which the oligomeric structure is clearly experimentally verified 2) include only crystal entries that fulfill a series of quality check criteria (see Methods), 3) focus on the range of interface areas where it is really difficult to distinguish crystal from biological contacts. We compiled two Duarte-Capitani datasets: one of large crystal contacts (DCxtal), the other of small biological interfaces (DCbio). DCxtal contains 78 entries validated as monomers, with 82 crystal interfaces of at least 1000 $Å^2$. For comparison, in the Bahadur set the lower limit for crystal interface area was set at 400 $Å^2$. Surely the growth in the number and average quality of available crystal structures has made the compilation of a sizeable dataset of large crystal contacts easier than in the past. DCbio consists of 74 oligomers, with 83 validated biological interfaces. Both datasets are listed in detail in Additional file 1: Tables S1 and S2, respectively. In Figure 1 we plotted the area distribution of the entries in our datasets and two others for comparison: the Ponstingl 2003 dataset of monomers and dimers [18] and the Bahadur homodimer and monomer datasets [22]. The boxplots show clearly very different area distributions, being our datasets a mixture of biological and crystal interfaces belonging exclusively to the overlapping area region. We also compared the DC sets with the PiQSi database [23]: while only 23 DC entries (out of 152) were present in PiQSi, their assignments were 100% in agreement with the PiQSi ones.

Geometry criterion: core size

The idea of dividing the residues of the interface, *i.e.* those that bury some surface area, into different classes appeared early in the protein interface literature. LoConte *et al.* [24] proposed a first classification based on atoms rather than residues, dividing them into 3 classes which they called A, B and C. The fully buried atoms formed class B, while classes A and C were subdivisions of the partially buried ones. Later Chakrabarti and Janin [25] introduced the concept of core residues as those residues having at least one fully buried atom. This definition was later used by Guharoy & Chakrabarti in their pioneering work on the relative average entropies of core and rim residues in interfaces [8]. Schärer *et al.* [15] substantially modified the definition of core residue, basing it on the percentage of the accessible surface area (ASA) that becomes buried upon interface formation. The cut-off for defining a residue as core was set by Schärer *et al.* at 95% burial (BSA/ASA). Levy [26] used a more complex scheme with 3 categories: core, rim and support. The scheme uses, as well as BSA and ASA, the relative surface accessibilities (rASA), i.e. the ASA of a residue X relative to its ASA in a reference extended tripeptide GLY-X-GLY. In Levy's definition a

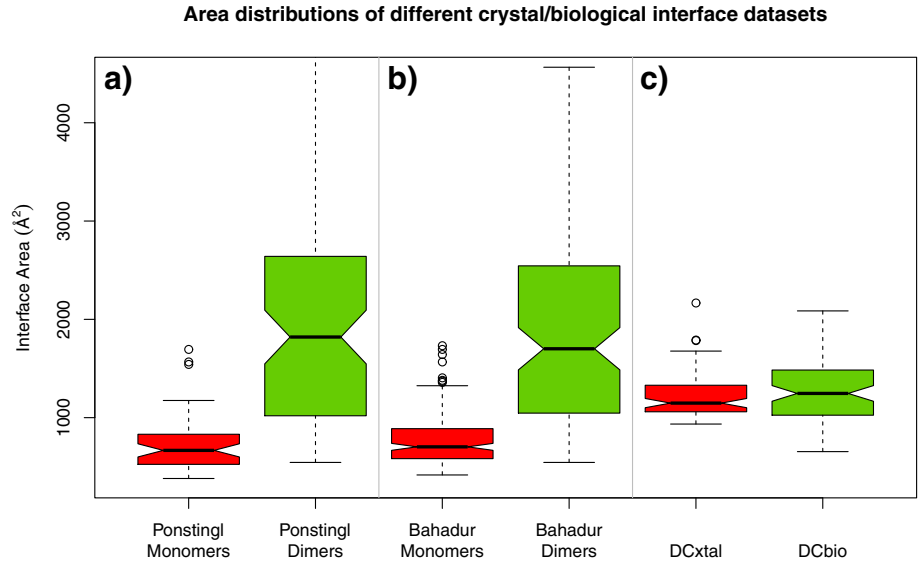


Figure 1 Distribution of interface areas in benchmarking datasets: Boxplots for a) Ponstingl Monomers in red and Ponstingl Dimers in green, b) Bahadur Monomers in red and Dimers in green c) DCxtal in red and DCbio in green. Our datasets focus on the range of areas where the two types of interfaces overlap the most, making it most difficult to predict their character.

residue is "core" if: 1) its rASA in the monomer is larger than 25% and 2) its rASA in the complex is smaller than 25% (Table 1). The Schärer definition proved effective when employed to divide interfaces into a rim and a core set, the average *Ka/Ks* ratios of which were then compared to classify the interfaces into "crystal" or "bio". As part of the present work we analyzed the predictive power of the three residue-based core definitions (Chakrabarti, Levy, Schärer) when using the number of core residues as a simple geometric criterion to categorize interfaces as biological or crystal contacts. Figure 2 displays the number of core residues (from now on called core size) found in our datasets of biological interfaces and crystal contacts by using the Chakrabarti, Levy and Schärer core definitions, respectively. The core size of each interface is plotted versus its area. Notably, while the two former core definitions lead to a quite strong correlation of core size with interface area

Table 1 Interface core definitions from the literature

	Chakrabarti	*Levy*	*Schärer*
Core definitions	Residues with at least 1 atom with BSA/ASA(u)=1	Residues with rASA(u)>0.25 & rASA(c)<0.25	Residues with BSA/ASA(u)> 0.95
Example bio interface: [PDB:1N8P] interface 1 (total BSA=1969 Å^2)	26+27	24+24	4+3

Interface residues are those for which BSA>0, core residues are then a subset of those. The values of core residue sizes for a typical biological interface example chosen from one of the entries in DCbio are shown (the 2 numbers corresponding to first and second partner of the interface). BSA is defined as BSA=ASA(u)-ASA(c), relative ASA as rASA=ASA/ASA(GLY-X-GLY). *u* and *c* stand for uncomplexed and complexed respectively.

(Pearson correlation coefficients 82% and 78%), the latter is much less correlated (Pearson 33%). Moreover in many cases it seems to clearly separate crystal from biological interfaces. For our two datasets it is able to tell bio interfaces apart from crystal interfaces with 80% sensitivity and 73% specificity, which makes it *per se* a powerful discriminator of interface character. In their 2004 work Bahadur *et al.* [22] presented two geometric parameters that were also very good at discriminating interfaces, namely the fraction of buried atoms and the non-polar interface area. It must again be underlined that the data used in that study was very different: their crystal interface areas were above 400 Å^2 whilst here our DCxtal interfaces are above 1000 Å^2.

As another way of displaying the predicting power of Schärer's core definition we produced ROC curves (Figure 3) depicting the ability of various geometrical parameters to predict the character of a) our DC bio/crystal interfaces and b) Ponstingl's bio/crystal ones. Schärer's definition outperforms the others and also the interface area as predictors. This difference only becomes apparent when using the datasets that focus on the difficult to predict region (a). If we use a more conventional dataset with a typical area distribution (b) the difference does not appear. This is striking as previous studies [9,27] of several geometrical interface parameters, including some based in Voronoi tessellation, found that area ranked first in prediction power compared to the other parameters.

Schärer's definition uses a percent burial cut-off to assign residues to core, so the question arises as to what an optimal value for this cut-off in terms of interface classification

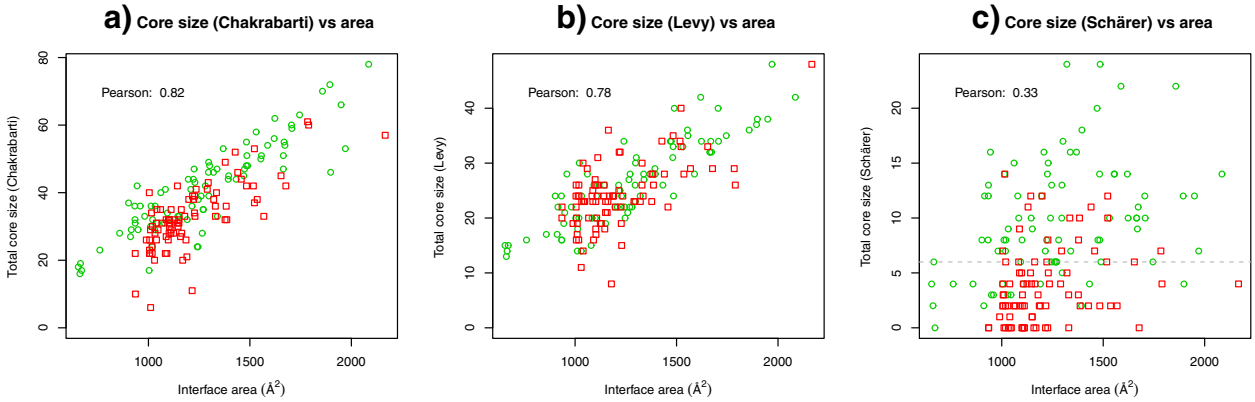


Figure 2 Correlation of core size in different definitions to area. Dots represent interfaces of the DCbio (green circles) and DCxtal (red squares) datasets. The first two definitions show high correlation, whilst the definition of Schärer, used in this work, has a much lower correlation. In the third plot we marked the core size value of 6 (cut-off used for the geometry classifier) with a horizontal line.

is. Strikingly, the 95% cut-off appears much more powerful than lower ones. We plot in Figure 4 the ROC curves of the core size at different cut-offs (95%, 50% and 10%) as predictors of interface character for our DC datasets. It is apparent that when one includes more and more partially buried residues the predictive power decays rapidly.

The core residues thus defined seem to be an essential interface determinant. Interestingly the definition is in agreement with that of hot spot residues introduced by Bogan *et al.* [28] and offers a possible explanation as to why the number of core residues is so powerful in distinguishing biological interfaces. In that study, the authors compiled a set of site-directed mutagenesis studies on interface residues and found that only a few well-buried residues contributed the most to the binding energy of the interface. Moreover, all residues that contributed significantly to the binding energy were fully (or nearly fully) buried, whilst partially buried residues were never found to significantly contribute to the energy. Thus, full burial was a necessary but not sufficient condition for a residue to be a hot spot.

It must also be noted that crystallographic accuracy is essential for the effectiveness of the geometry criterion, the full power of which can only be seen when using sets of good quality crystal protein structures. A striking example is the structure of bovine interferon gamma, solved first at 3 Å resolution ([PDB:1RFB]) and later again at 2 Å resolution ([PDB:1D9C]) in the same crystal form. The area of the dimer interface changes from 2600 to 3600 Å^2 from the first to the second, but more

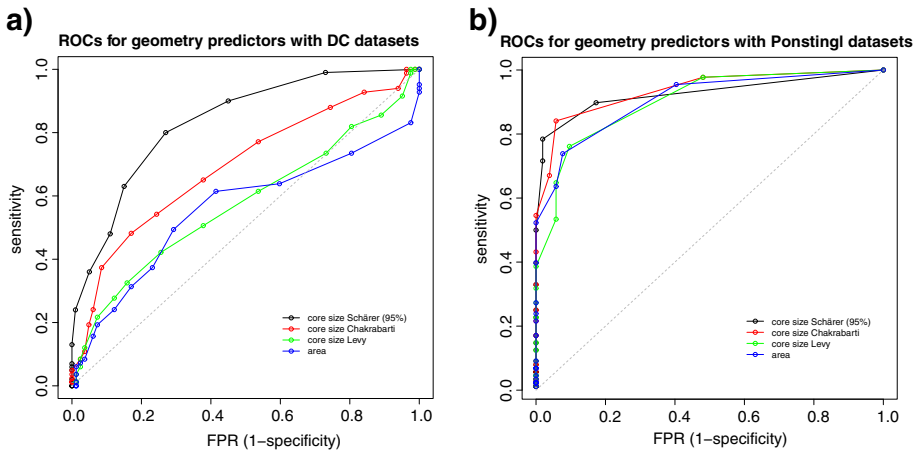


Figure 3 ROCs for different geometric indicators. The ROC curves represent the predictive power of different geometric parameters: core size (Schärer's definition), core size (Chakrabarti's definition), core size (Levy's definition) and total buried surface area. In panel a) the datasets used are our DCbio/DCxtal, whilst in panel b) the Ponstingl datasets were used. Not much difference can be appreciated if using Ponstingl's dataset, since it contains interfaces that are too clearly separable by area. When we use the DC datasets, it becomes apparent that Schärer's core definition exhibits superior performance compared to the other geometric indicators.

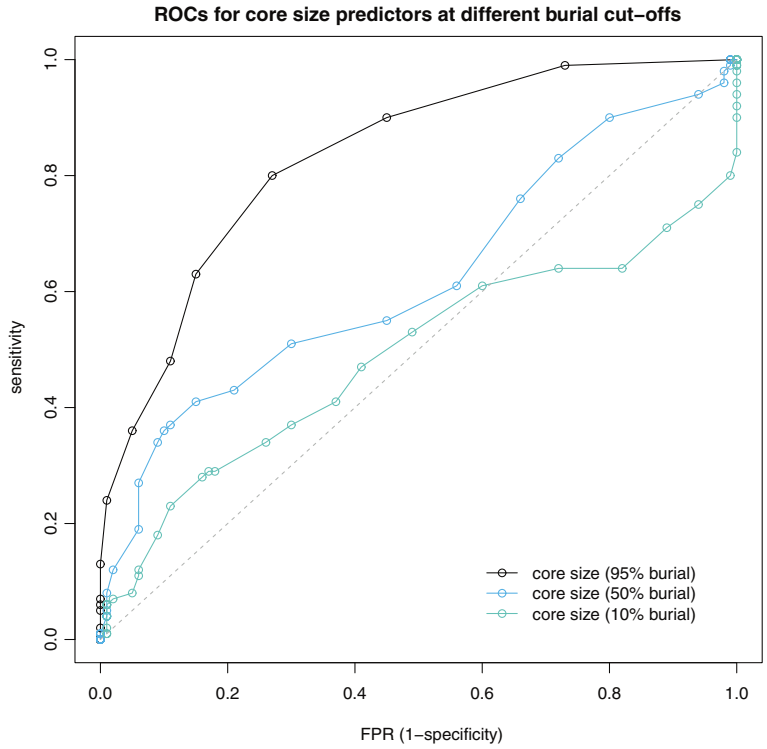


Figure 4 Schärer's core definition at different cut-offs. ROC curves for Schärer's core size at different BSA/ASA cut-offs as predictor for the DC datasets. The 95% burial cut-off has a clear advantage over the lower cut-off core definitions.

importantly the number of core residues leaps from 1 in the first case to 36 in the second.

Estimation of selection pressure: sequence entropies

As mentioned above, in this work we decided to move from the *Ka/Ks* ratio selection pressure metrics to sequence entropies at the amino-acid level. We realized that we could see very good differential selection pressure signal at the interfaces by carefully choosing the homolog sequences to measure the entropies. Most importantly we decided to be very conservative in the amount of homologs to use, cutting the homolog list at a sequence identity value as high as 60% (extending to a hard cut-off of 50% when not enough homologs are found). There are mainly two reasons for this choice.

First, by staying in the very high identity region we avoid the risk of introducing errors in the alignments and we can rely on the assumption that the structures of homologs used in the alignment are very well conserved. From knowledge gathered over the years of CASP structure prediction experiments, it is known that alignment accuracy is very good only down to ~50% sequence identity, medium to good in the 30-50% identity region and low below 30% identity (the "twilight zone") [29,30]. These assessments done over the different CASP experiments are based on the gold-standard of a structural alignment to the best template [29].

As a second point the quaternary structure of proteins and thus interfaces seem to be less conserved than that of the tertiary structure. Poupon and Janin [31] estimate that 40% is a reasonable limit to the reliability of a good quaternary structure homology, thus it seems dangerous to consider sequence homologs below that 40% level.

Strikingly, almost all methods for interface classification or prediction until now have used much lower sequence identities for measuring conservation. For instance a few studies [7,8,32,33] used the well-known HSSP database to get their alignments. HSSP uses 25% as the identity cut-off for sequences with length above 80 residues (a majority of PDB proteins these days) [34]. Valdar *et al.* [6] select their homologs by performing a maximum of 20 PSI-blast iterations with an inclusion e-value cut-off of 10^{-40} which results in identities as low as 5%. Caffrey *et al.* [35] even compared two types of alignments: a "diverse" one, aimed at capturing paralogs, and a "close" one to contain only orthologs. The former had a very generous homolog inclusion cut-off (blast with e-value cut-off of 0.001) while the latter took close orthologs from selected species in the same taxonomic kingdom. This second type of alignment, although more stringent, is not comparable to those computed here, as it typically contained very few sequences (10 to 15).

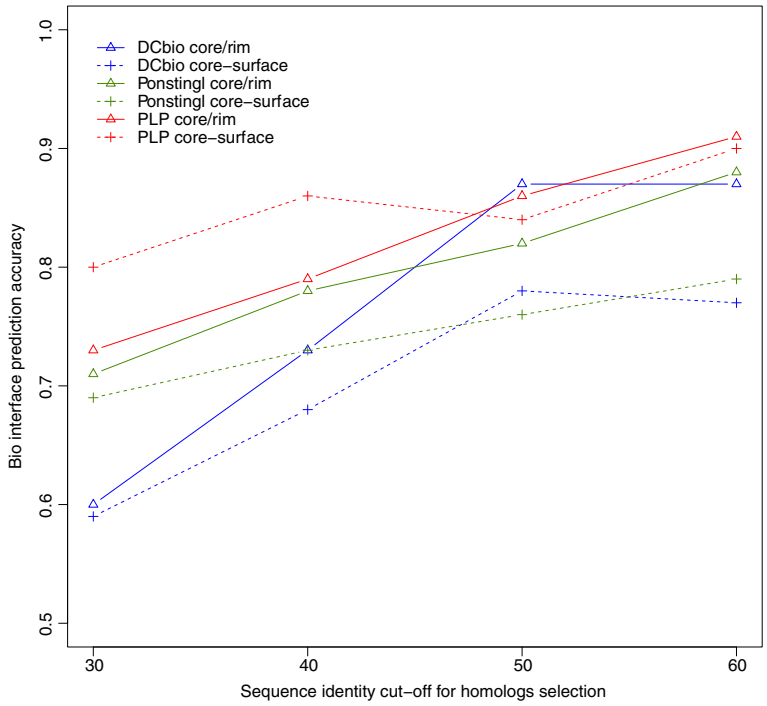


Figure 5 Our prediction accuracies on biological interfaces versus identity cut-offs used for homolog selection. The prediction accuracies of our 2 evolutionary methods (core-rim entropy ratio with solid lines and core-surface entropy score with dashed lines) is plotted against different identity cut-offs for selection of homologs to be included in the alignments. For all datasets accuracies are lower when more distant homologs are used in the alignments.

In order to see how the choice of identity cut-off affects our interface predictions we studied the accuracies of predictions with variable identity cut-offs. The results are presented in Figure 5. As we lower the sequence identity cut-off for inclusion of homologs in our alignments the accuracy of the evolutionary predictions clearly degrades. The behavior was similar across different sets of biological interfaces datasets (DCbio, Ponstingl dimers and PLP enzymes). We achieved optimal results with a combination of 60% soft identity cut-off and 50% hard identity cut-off (see Methods).

Choosing a more stringent cut-off is only possible thanks to the size that sequence databases have reached in the last few years. As the growth will only continue in the foreseeable future we believe that our conservative approach will continue giving the best signal to noise ratio in measuring differential selection pressure of interfaces.

Core versus surface scores

One of the earliest attempts to use evolution to predict biological interfaces [7] compared average sequence entropies of interface residues versus those of the other surface residues. As discussed in the Introduction, this approach was hampered by bias caused by patches of low-entropy surface residues corresponding for instance to binding sites or external active sites. In the search for an additional evolutionary prediction criterion, we took inspiration from that early attempt and introduced an approach comparing the average sequence entropies of interface core residues and of surface residues. In order to reduce bias in the calculations, we employ random pooling of surface residues. Given an interface with N core residues, we sample random pools of N surface residues so that we then can compare the entropy of the core residues versus that of the distribution of surface samples. We then give the final score as the distance of the average core entropy to the mean of the surface samples in units of their standard deviation, in a Z-score-like approach.

Core-surface scores provide a measure of the selection pressure acting on the key residues of an interface compared to a surface "baseline" estimated from the randomly pooled surface residues. In order to further reduce bias, only those surface residues that are involved in none of the interfaces found in the crystal are used for pooling.

Valdar and Thornton [17] did also employ a surface sampling approach in analyzing a limited set of homodimer interfaces, though in their case the statistical significance of the interface versus surface conservation was assessed *via* P-values. Later, Caffrey *et al.* [35] followed

that approach and used a Z-test for significance estimation, but concluded that the measured evolutionary signal was not sufficient to predict interface patches from conservation information alone.

Combining information from the different criteria

As described above we employ three different indicators of the interface character: a geometric one and two evolutionary, core versus rim entropy ratio and core versus surface entropy score. To offer a final prediction of interface character we set out to combine the different indicators into a single call. We decided for a simple majority voting system, where we place more confidence on the evolutionary calls (see Methods). In the case that not enough suitable homologs are available for a certain protein structure, making it impossible to employ the evolution-based criteria, the final call is based on geometry only. In addition we employed the results from the compilation of the DCxtal contact dataset (see Methods for details) to establish hard limits for biological or crystal contact character: areas above 2200 $Å^2$ are always considered biological, while areas below 400 $Å^2$ are always considered crystal, irrespective of the other indicators. The low hard area limit criterium refers to non-induced [36] protein-protein interfaces, and does not apply to protein-peptide ones.

Engineering artifacts in the PDB: a word of caution

A further novelty we introduced in our interface classifier method is that of checking for engineering artifacts in the structure being analyzed. This important aspect is, to our knowledge, mostly neglected by computational methods attempting to classify crystal interfaces. In order to produce, characterize and crystallize proteins, structural biologists often need to introduce modifications into their wild-type sequences. These range from point mutations to insertion of affinity tags or to total chimeric constructs. We deal with this issue by first of all finding a reference UniProt sequence for the given PDB sequence. Multiple UniProt assignments to a single PDB entry usually indicate a chimeric construct (*e.g.* the recent structure of the channelrhodopsin light-gated cation channel [37][PDB:3ug9]. In that case, no evolutionary prediction is run and interface classification relies on core size only. If a reasonable reference UniProt alignment exists (as defined by sequence identity and coverage thresholds) then we attempt to predict the interface with all three criteria. In these cases we further check whether the core and rim residues to be scored locate in a region that aligns properly to the reference. Warnings are produced for mismatches; if the number of mismatches exceeds a threshold, again no evolutionary prediction is carried out and the final call is geometry-based.

Parameter optimization and performance

Several parameters are used in classifying an interface as biological or crystal. Especially important are the cut-offs used for each of the scores: core size (geometric indicator), core versus rim entropy ratio and core versus surface entropy score. In order to optimize those we used our manually annotated DCxtal and DCbio datasets, which contain entries with experimentally verified quaternary structure assignment and with areas in the difficult range 1000–2000 $Å^2$. The optimization process with these datasets led to the following cut-off values: 6 core residues for geometry, 0.75 for entropy core/rim ratio and −1.0 for core versus surface scores.

Finally, in order to benchmark our method with a separate independent set we used the well-known Ponstingl 2003 [18] sets of monomers and dimers which we minimally modified (see Methods). This dataset has the advantage of having been employed several times as benchmark in the literature [9,12,38]. In the case of PISA [36] it was also used as optimization set.

In Table 2 we present the results of the optimization and benchmarking steps and for reference we include the PISA performance on the same sets (see Methods for details on how the PISA performance was measured). The three different methods are first assessed separately and then as a single combined predictor. Additionally to the two datasets DC and Ponstingl we also include the statistics for the Bahadur set (a superset of Ponstingl's) for completeness. Overall, the performance of our final combined predictor compares favorably to that of the PISA server in the 3 sets. The geometric predictor alone is able to classify the interfaces with high accuracy and is helped by the evolutionary ones to further improve the performance in the final call. It must be noted that the evolutionary predictors are in themselves very powerful at classifying interfaces, with sensitivity/specificity figures ranging from 64% to 87%. These figures are not directly comparable to those of the geometric predictor or the combined predictor as they are based on the subset of entries that could be predicted at all (prerequisites are that at least 10 homologs are available and that enough core/rim/surface residues exist). In the analysis of wrong evolutionary predictions we often found cases with problematic alignments, *e.g.* with inhomogenous sequence identity distribution of homologs. Viral or archaeal proteins seem particularly prone to this kind of problem. We are convinced that better filtering and selection criteria will help in further improving the performance of the evolutionary predictors.

Performance with sequence data growth

In order to more precisely assess the performance of our method we studied the behavior of the evolutionary predictions with the change in sequence data over the last

Table 2 Classification statistics

	EPPIC (based on UniProt 2012_10)											
	# entries		**Geometry**		**Entropy core-rim**		**Entropy core-surface**		**Combined**			
	Bio	**Xtal**	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Sens.**	**Spec.**	**Acc.**	**Sens.**	**Spec.**	**MCC**
DC (optimization)	83	82	0.80	0.73	0.82(68)	0.66(64)	0.87(69)	0.76(67)	**0.81**	**0.88**	**0.73**	**0.62**
Ponstingl (benchmarking)	88	52	0.85	0.92	0.84(76)	0.66(29)	0.85(75)	0.79(29)	**0.89**	**0.90**	**0.87**	**0.76**
Bahadur (benchmarking)	121	185	0.88	0.88	0.82(103)	0.64(114)	0.86(104)	0.77(114)	**0.86**	**0.89**	**0.84**	**0.72**
	PISA											
	Acc.	**Sens.**	**Spec.**	**MCC**								
DC (optimization)	0.79	0.95	0.63	0.62								
Ponstingl (benchmarking)	0.84	0.89	0.77	0.66								
Bahadur (benchmarking)	0.77	0.89	0.69	0.57								

Classification statistics for our own compiled datasets ("DC"), composed of DCxtal (crystal interfaces) and DCbio (biological interfaces), for the Ponstingl 2003 dataset of monomers (crystal interfaces) and dimers (biological interfaces) and for the Bahadur datasets (monomer and dimers). We first present the statistics for each of our indicators separately and the statistics for the combined predictor. PISA statistics compiled by us are shown in a separate table. Statistics are given in terms of sensitivity or rate of correct biological interface predictions and specificity or rate of correct crystal interface predictions. The statistics for the evolutionary methods are based on the total number of interfaces that could be predicted (enough homologs and enough core/rim/surface residues). The numbers for each case are indicated in parentheses together with the corresponding sensitivity or specificity. As well as accuracy values we present the Matthews correlation coefficient (MCC) which gives a better assessment of the predictions in cases where the positive and negative sets are unbalanced (as is the case with the Ponstingl sets). All EPPIC evolutionary predictions are based on UniProt release 2012_10.

years. The UniProt database has seen an exponential growth aided mainly by an improvement in sequencing technologies that even outperforms Moore's law [39]. We studied the performance dependence of our interface evolutionary predictions with the growth of sequence databases by using archived UniProt versions from the first release appeared in December 2003 to the current one almost 10 years later. The first and more important effect that we observed is a dramatic increase in prediction coverage as the UniProt database grows. For the Ponstingl datasets, coverage rose from 27% in 2003 to 65% in 2012. We are able to predict a particular entry whenever we can find at least 10 non-redundant sequence homologs within 50% identity of the query. Additionally we tried to assess whether the accuracy of the scores increases as alignments get enriched with more sequence data. We thus studied the evolution of the core-surface scores in biological interfaces from a few datasets (DCbio, Ponstingl dimers and PLP enzymes), plotted in Figure 6a. The score distributions across all interfaces exhibit a downwards trend both in terms of median scores and of their spread. Contrastingly Figure 6b present the scores across time for crystal interfaces (DCxtal and Ponstingl monomers), where a slight upwards trend can be observed and not much variation in the spread.

Web server

In order to make the EPPIC approach easily accessible to the structural biology and bioinformatics community, we built a web server (http://www.eppic-web.org), with a front-end design centered on clarity and usability. To that end we created a rich interactive web application, based on the Ext-GWT (http://www.sencha.com/products/extgwt) framework. As a minimum input, the user has simply to provide the PDB code of the entry to be analyzed or to upload a coordinate file in PDB or mmCIF format. The user can also access an “Advanced” input panel that allows for changing the parameters for homolog selection and alignment. A collapsible panel on the left provides an overview of the currently running and of the completed jobs. The results page (Figure 7) consists of a top panel, showing the key information about the job and of a dynamic table listing all interfaces present in the crystal lattice. Each row of the table corresponds to an interface, represented as a clickable cartoon-style thumbnail, and shows additional information about the interface. The last columns give the prediction calls (bio or xtal) for all three approaches (geometry, entropy core-rim ratios, entropy core-surface scores) and the final combined call. As an optional extra column, warnings are shown if the residues involved in the interface do not properly align to the reference UniProt entry or other kinds of issues are found in the interface geometry. By clicking on an interface thumbnail, the user can access a 3D view of the interface itself, either through JMol [40] (browser-based, no need for an installed viewer), a local molecular viewer (PDB file) or a PyMOL [41] pse session file.

A practical example

An example of using EPPIC in the context of an important structure biology problem is provided by the work of Zhang *et al.* [42] on the mechanism of activation of epidermal growth factor receptor (EGFR), which is based on dimerization. The authors determined the

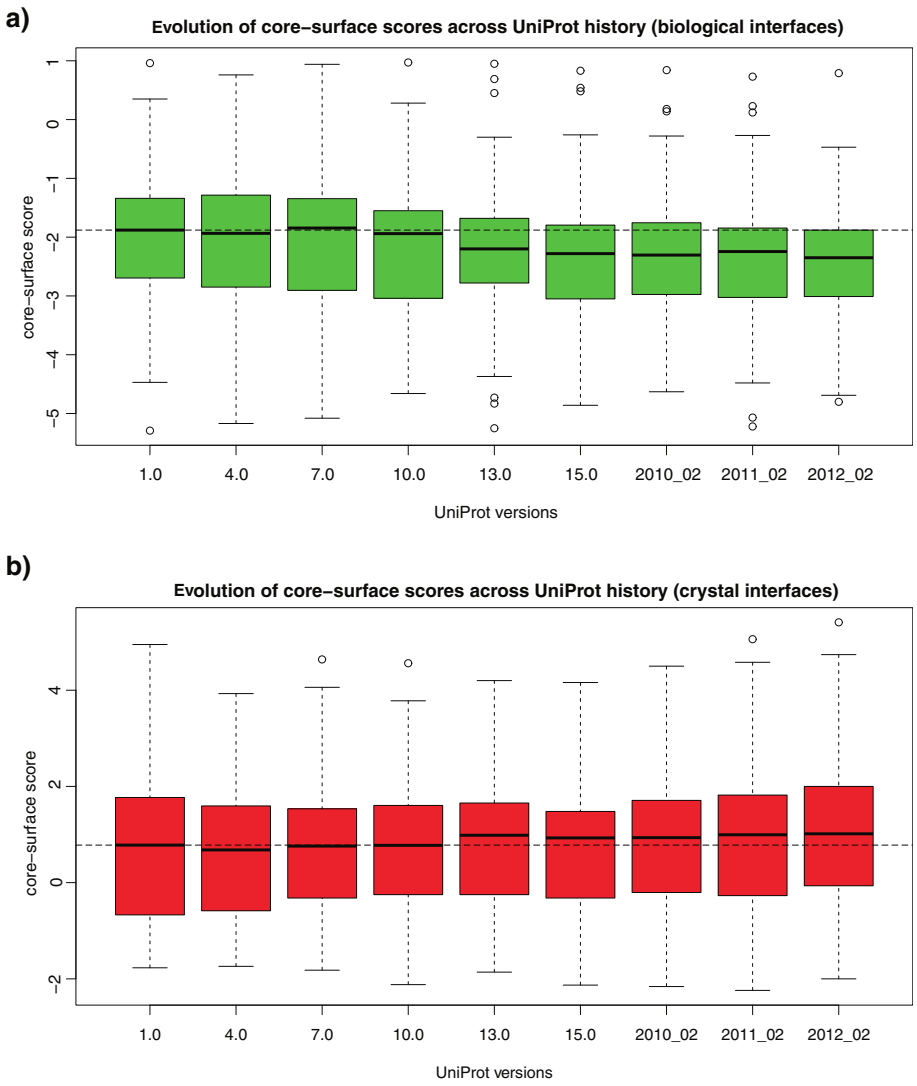


Figure 6 Core-surface score variation across UniProt history. The core-surface scores improve on average as more sequence data has become available. Plotted are core-surface scores of **a)** biological interfaces (from DCbio, PonstingI Dimer and PLP datasets) and **b)** crystal interfaces (from DCxtal and PonstingI Monomer datasets). The lower the score the stronger the indication of biological interface (our cut-off for classifying bio/crystal is set at −1). The median score for UniProt version 1.0 (2003) is denoted by a dashed line. The chosen versions are separated in time by approximately one year.

crystal structure of the EGFR kinase domain ([PDB:2GS2]), where either a symmetric or asymmetric dimer, of very similar size (950 and 990 A^2, respectively), can be chosen as the biologically relevant entity mediating activation. A symmetric dimer, already determined by Stamos *et al.* [43] in a different crystal form, was computationally analyzed by Landau *et al.* [44], who proposed it, among six possible dimer choices, as the key contact controlling inactivation of the receptor. Zhang *et al.* settled the issue with a series of mutagenesis experiments that identified the asymmetric dimer as the relevant one. EPPIC analysis of entry [PDB:2GS2] clearly indicates the Zhang asymmetric dimer as biologically relevant and the symmetric one as a crystal contact. It does so based on clear signals by the entropy core-rim and core-surface criteria, which lead to a correct call for this difficult case in which both interfaces exhibit similar geometrical features (similar number of core residues). Strictly speaking, the asymmetric dimer is unviable since such heterologous interfaces can extend to infinite fibers [45]. Zhang *et al.* do acknowledge this issue and attribute the apparent contradiction to the fact that the crystallized construct is only an intracellular fragment of the full length membrane protein. The symmetric and asymmetric dimers of [PDB:2GS2] are shown in Figure 8

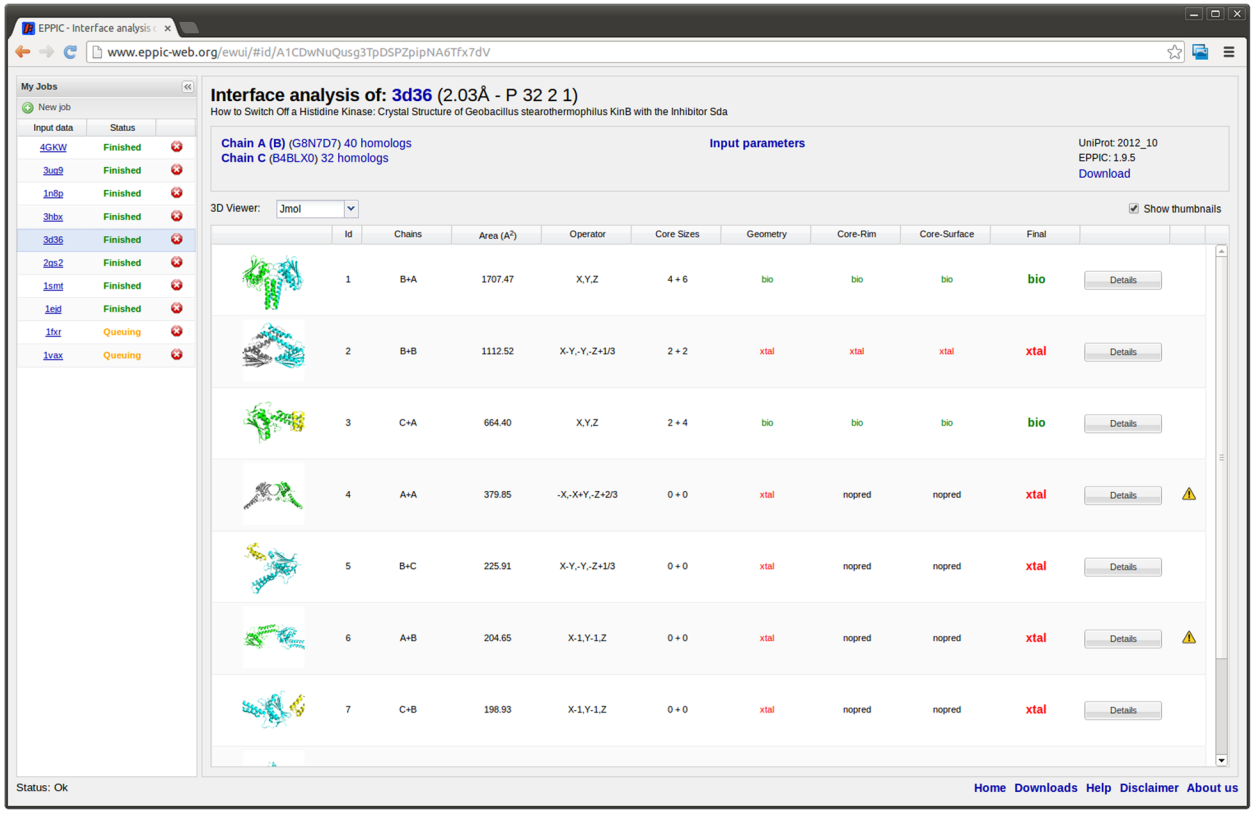


Figure 7 Typical output display of the EPPIC server.

(panels a) and b), respectively) as they would appear to the user in the respective PyMol pse session files provided by the EPPIC web front-end.

Conclusions

We present here a new, highly effective and easy-to-use method addressing an important issue in structural biology and bioinformatics: that of distinguishing crystal contacts from biologically relevant interfaces. The importance and spread of this problem is now widely recognized: as an effective method to solve it, EPPIC will significantly help in the interpretation of crystal structures, in guiding biochemical experiments on protein-protein interfaces and hybrid approaches in which single components solved by crystallography are to be assembled into large supramolecular entities. Two important conclusions can be drawn from this study: first, that fully buried residues are a key determinant of biological protein-protein interfaces; second, that a stringent sequence selection for the multiple sequence alignments used to measure the evolutionary signal provides a more robust and less noisy way to detect the footprint of evolution in interfaces. This is especially important as the incessant growth of sequence databases fueled by new high-throughput technologies will only increase the usefulness of evolution-based methods. EPPIC bears significant potential for further developments, the most straightforward one being automatic inference of quaternary structure assemblies from the interface predictions, thus providing a complete pipeline from crystal structures to putative biological assemblies. The method is applicable to many problems in both structural biology and structural bioinformatics, to name just a few: validation of structures of oligomeric proteins and of protein complexes, detection of crystal contacts in which one of two partners mimics a biological partner, prediction of protein-protein binding sites in the absence of the structure of a complex and the validation of models of complexes and oligomers.

Methods

Compilation and annotation of new reference datasets

In order to compile our monomer dataset (DCxtal) we first gathered a subset of PDB entries by using the advanced query feature of the RCSB PDB site (http://www.pdb.org) on the 21st of December 2010 with following parameters: 1 chain in the biological unit (biounit), resolution better than 1.8 Å, Rfree below 30%, Rsym below 10% and with a sequence redundancy filter at 90% identity. We then calculated all possible interfaces for

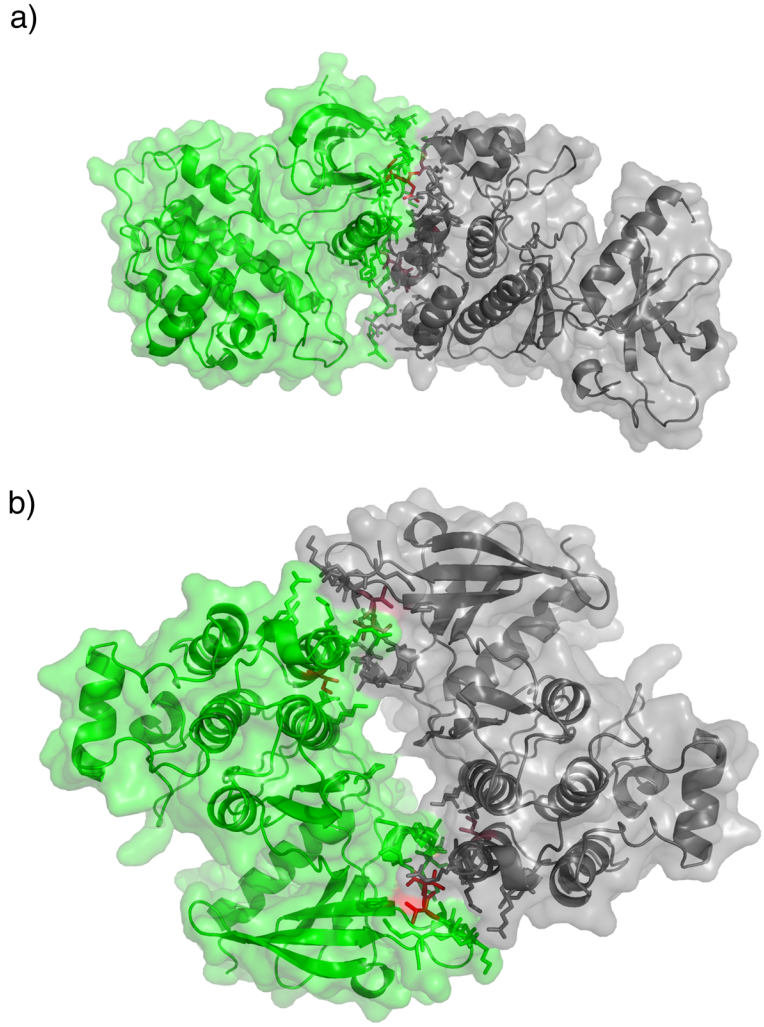


Figure 8 Identifying the biologically relevant interface of the EGFR kinase. Asymmetric (top) and symmetric (bottom) dimers in the structure of the epidermal growth factor receptor kinase ([PDB:2GS2]). The two interfaces appear as in the respective PyMOL pse sessions downloadable from the EPPIC web front-end by clicking on interface thumbnails (surface rendering was added for clarity).

those entries, taking for our further manual curation those with an interface area above 1000 Å^2. A further quality control eliminated those entries that generated more than 5 clashes (atoms within 1.5 Å) between chains during the interface calculation process. With this procedure we aimed at finding all putative large crystal interfaces from crystal structures of good crystallographic quality in the PDB. This filtering resulted in a set of 378 PDB entries, which we manually curated by looking into their main references and other literature when necessary. We only took an entry as a candidate for our DCxtal dataset when clear experimental evidence for monomeric state was found in the literature, usually provided by size exclusion chromatography, analytical ultracentrifugation or light scattering techniques [31]. We discarded entries with dubious features or experimental evidence: for instance, putative domain swaps (by visual inspection), debated oligomeric state with conflicting experimental data in the literature or cases where experimental evidence referred to a different fragment than the crystallized construct.

For the DCbio dataset we first took entries that in the above procedure were found to be clearly experimentally verified to be multimeric (thus mostly annotation errors in the PDB as we initially selected entries with 1 chain in the biounit). Then we added 10 PLP enzymes with biological interfaces with areas below 2000 Å^2. PLP enzymes are known to exist always as dimeric or higher oligomeric assemblies [46]. Finally we proceeded with a similar methodology as above by filtering the PDB for good quality structures with 2 chains in the biounit, aiming to find putative dimers. The interfaces for them

were calculated and those with areas between 900 and 1400 Å^2 were chosen for manual curation by literature search as above.

Hard area limits

In the above annotation effort, 41 entries contained extremely large (>2000 Å^2) putative crystal interfaces, from which we could only validate one real monomer ([PDB:1LF2], with an area of 2171 Å^2). All others were either errors in the biological unit annotations or dubious cases. On this basis we set a hard area cut-off of 2200 Å^2, above which interfaces are directly called biological without considering the other indicators. We could count only 130 other putative monomers in the PDB (December 2010) having their largest interface area above 2200 Å^2, resolution <3.0 Å, Rfree<35% and fewer than 5 clashes. Thus, our sample of 41 manually curated monomers represents about a quarter of all putative large monomer interfaces with reasonable quality in the PDB, so the chosen hard cut-off can be considered significant.

Interface calculation and geometry criterion

We calculated the interfaces for a given entry by applying all symmetry operators corresponding to the entry's space group and finding any pair of chains that had at least 1 atom of each side within a distance of 6 Å. We implemented the interface calculation in Java and integrated it in our code. For surface calculations, we used the implementation of the Shrake and Rupley algorithm [47] written by Bosco Ho (http://boscoh.com/protein/asapy) which we ported into Java. A ball radius of 1.4 Å was used to calculate the Accessible Surface Areas (ASA). Surface residues were considered those exposing more than 5 Å^2 of their surface. We computed both the ASA of complexed and uncomplexed subunits, finding by subtraction the Buried Surface Area value (BSA):

$$BSA = \frac{1}{2}\left(ASA_{uncomplexed} - ASA_{complexed}\right)$$

Surface residues with BSA>0 constituted the interface. We then followed Schärer's [15] definition to assign the core residues as those with BSA/ASA$_{\text{uncomplexed}}$>0.95. Interfaces with more than 6 core residues were considered biological. This value was found in an optimization procedure carried out on the DCxtal and DCbio datasets that maximised both sensitivity and specificity.

Evolutionary scoring

To calculate sequence entropies for each of the residues of a given PDB structure we used the following procedure: 1) We found the reference UniProt identifier for the PDB sequence by using SIFTS (http://www.ebi.ac.uk/pdbe/docs/sifts) or blasting, in order to control for possible engineering performed on the PDB sequence. 2) Using the reference UniProt sequence we searched the UniRef100 (http://www.ebi.ac.uk/uniref/) database through BLAST [48] to find putative homologs. Only the matching PDB subsequence of the UniProt reference was used for the BLAST search. 3) We then applied sequence identity (soft cut-off of 60% identity, relaxing in 5% steps down to 50% identity until at least 10 homologs were found) and coverage (80%) filters and a hard maximum number of sequences of 100. 4) We then clustered the sequences by using BLASTCLUST [48] and choosing a single representative from each cluster. We did this in an iterative way by starting with a 98% identity clusters and reducing stepwise this threshold if more sequences needed to be eliminated to reach the hard maximum of 100 sequences 5) We finally used the CLUSTALO [49] program to perform a multiple sequence alignment of the selected homologs. 6) Based on that sequence alignment sequence entropies were calculated. The Shannon entropy of an alignment column i is given by:

$$s(i) = -\sum_k p_i(k)\log(p_i(k))$$

where $p_i(k)$ is the probability of a residue of class k being at position i of the alignment. We used a reduced amino-acid alphabet with 10 amino acid classes as proposed by Murphy *et al.* [50].

Entropy values were finally mapped back to the PDB sequences, so that we could compute from those core versus rim ratios and core versus surface scores as described above. For entropy scoring the core residues were chosen with a more relaxed criterium of 70% burial (assigning the remaining interface residues as rim) in order to achieve more statistically significant comparisons. Only if more than 8 of them exist above 70% burial, we make an evolutionary prediction. For the core versus surface scores calculation we drew 10000 samples of N residues (N being the number of core residues in the analysed interface) from surface residues belonging to none of the interfaces found in the crystal.

Combined predictor

The combined predictor is based on a simple consensus vote from the 3 methods: geometry, entropy core over rim ratio and entropy core versus surface scores. The majority vote (2 out of 3) of the separate calls gives the final prediction. If an evolutionary prediction cannot be made, due to lack of enough homologs or to an insufficient number of core residues, then the final call is the geometric one. In some cases one of the two evolutionary measures fails. For instance core over rim ratio can fail if too many of the rim residues are mutated, or the core-surface prediction can fail if the surface from which

to draw residues is too small. In such cases if the geometry and evolution call do not agree, preference is given to the evolutionary one. Additionally hard area limits are used as described above. Special cases like interfaces with disulfide bridges in wild-type residues are treated as biological, disregarding the other indicators.

Optimization and benchmarking

We optimized the different parameters against the DCxtal and DCbio datasets. A pooled dataset using both sets of biological and crystal contacts was created and used for the runs. A True Positive was counted when our classifier was able to assign a biological interface as biological, True Negative when it could assign a crystal interface as crystal. We ran the predictions with several cut-off parameters for each of the three methods and chose the set of parameters that maximised accuracy ((TP+TN)/(P+N)). In the final statistics together with the accuracy we also quote the sensitivity (*i.e.* True Positive Rate or correct bio predictions from all possible biological interfaces) and the specificity value (*i.e.* True Negative Rate or correct xtal predictions from all possible crystal interfaces). PISA predictions were assessed as follows: for each entry in the pooled dataset the PISA interfaces and assemblies were downloaded as xml files (http://www.ebi.ac.uk/msd-srv/prot_int/pi_download.html). The first item in the assemblies list was taken as the PISA prediction. It was then checked whether the interface of interest was in the list of interfaces engaged by the assembly. The prediction for that interface was then assigned as biological. The interface was assigned as crystal if a) the interface of interest was not in the list of engaged interfaces b) no assembly prediction was given. If the PISA prediction fell in the "grey region of complex formation criteria" then it was considered as a failed prediction and not counted as either biological or crystal.

The Ponstingl 2003 dataset used here for benchmarking consists of two subsets: 1) a crystal interfaces set: largest interface from each entry in the Ponstingl monomers set; 2) a biological interfaces set: largest interface in each of the Ponstingl dimers set. We minimally modified the entries from the original version published in 2003 [21] to make sure the set was up to similar standards of accuracy as our own compiled sets. We did manual curation of 10% of its entries, finding in that process a few problems with the crystallographic quality of some entries and in some cases with the experimental oligomeric assignment. The entries that were modified were:

- in monomers dataset: removed [PDB:1A8O] and [PDB:2ABX] as they are known to be dimers, removed [PDB:2HEX] that is a debated monomer/decamer, see for instance discussion in Schärer *et al.* [15]
- in dimers dataset: entry [PDB:1RFB] (3Å resolution, no Rfree available) was replaced by [PDB:1D9C] (2Å resolution, Rfree 0.27)

Two additional datasets were used: Bahadur's monomer [22] and dimer datasets [51] in benchmarking and in Figure 1; and the PLP enzymes biological interfaces dataset from Schärer *et al.* [15] in Figures 5 and 6a.

Sequence data growth benchmarking

For the historical study section we first downloaded selected versions of the UniProt archived data available at ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases. We chose 9 versions that were distanced by approximately a year from each other and ranged from December 2003 to February 2012: 1.0, 4.0, 7.0, 10.0, 13.0, 15.0, 2010_02, 2011_02 and 2012_02. The interfaces used to study the score variation across time are further selected from the full lists by choosing only those that have a clear progression in the number of non-redundant homologs: between 5 and 15 homologs available in version 1.0 and more than 20 homologs available in version 2012_02.

Software

The core EPPIC code was written in Java using the OWL Java library for structural bioinformatics (http://www.bioinformatics.org/owl/) and is licensed under the GPL. The source code is available at the Subversion repository https://systemsx02.ethz.ch/svn/crk. All algorithms have been integrated in the Java code, including interface calculation and ASA calculations. Blast and Clustal Omega are the only external tools, which we then interfaced from Java. The UniProt JAPI (http://www.ebi.ac.uk/uniprot/remotingAPI/) is used for retrieving UniProt data. The web server is written in Java using the Ext-GWT framework (http://www.sencha.com/products/extgwt) and uses Hibernate (http://www.hibernate.org/) together with a backend MySQL database system for data persistency. The job scheduling in the computational backend is done through the open source Open Grid Scheduler/Grid Engine (http://gridscheduler.sourceforge.net/) system. A command-line version of the interface classification software is available for download at the web address http://www.eppic-web.org/downloads/eppic.zip. The web server is essentially a Web GUI to the command line program.

All plots were generated with R [52]. The PyMol [41] molecular graphics system was used for creating figures, thumbnails in server and extensively for analysis.

Additional file

Additional file 1: Tables S1 and S2. Manually curated monomer and oligomer DC datasets, with experimental evidence from the literature.

The "area" column refers to the largest interface in the protein crystal. References mostly given as PubMed id numbers linking to abstracts. Experimental evidence abbreviations used: SEC size exclusion chromatography; AUC analytical gel filtration; AUC (SV) analytical ultracentrifugation sedimentation velocity; SLS, DLS, LS (static/dynamic) light scattering; MALS multi-angle light scattering; MALLS multi-angle laser light scattering; CCL chemical cross-linking; FRET fluorescence resonance energy transfer; NMR nuclear magnetic resonance; SAXS small angle x-ray scattering; MS mass spectrometry; native-PAGE native polyacrylamide gel electrophoresis.

Abbreviations

EPPIC: Evolutionary Protein Protein Interface Classifier; PDB: Protein Data Bank; ASA: Accessible Surface Area; BSA: Buried Surface Area; DCxtal: Duarte-Capitani crystal interfaces dataset; DCbio: Duarte-Capitani biological interfaces dataset.

Competing interests

The authors declare no competing interests.

Authors' contributions

JMD, MAS and GC performed the analysis of the data. JMD and AS developed the software. JMD and GC conceived and designed the study and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Financial support from the Research Committee of the Paul Scherrer Institut to GC (grant numbers FK-05.08.1 and FK-04.09.1) is gratefully acknowledged, as is support from SyBIT (head Dr. Peter Kunszt).

Author details

[1]Paul Scherrer Institut, Villigen CH-5232, Switzerland. [2]SyBIT, ETH Zurich, Zurich, Switzerland. [3]Present address: Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich CH-8093, Switzerland.

Received: 2 August 2012 Accepted: 15 December 2012
Published: 22 December 2012

References

1. Leibundgut M, Jenni S, Frick C, Ban N: **Structural basis for substrate delivery by acyl carrier protein in the yeast fatty acid synthase.** *Science* 2007, **316:**288–90.
2. Huber EM, Basler M, Schwab R, Heinemeyer W, Kirk CJ, Groettrup M, Groll M: **Immuno- and constitutive proteasome crystal structures reveal differences in substrate and inhibitor specificity.** *Cell* 2012, **148:**727–38.
3. Bilokapic S, Schwartz TU: **3D ultrastructure of the nuclear pore complex.** *Curr Opin Cell Biol* 2012, **24:**86–91.
4. Hoelz A, Debler EW, Blobel G: **The structure of the nuclear pore complex.** *Annu Rev Biochem* 2011, **80:**613–43.
5. Janin J: **Specific versus non-specific contacts in protein crystals.** *Nat Struct Biol* 1997, **4:**973–4.
6. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *J Mol Biol* 2001, **313:**399–416.
7. Elcock AH, McCammon JA: **Identification of protein oligomerization states by analysis of interface conservation.** *Proc Natl Acad Sci USA* 2001, **98:**2990–4.
8. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces.** *Proc Natl Acad Sci USA* 2005, **102:**15447–52.
9. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A: **DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions.** *Bioinformatics (Oxford, England)* 2008, **24:**652–8.
10. Zhu H, Domingues F, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types.** *BMC Bioinforma* 2006, **7:**27.
11. Mitra P, Pal D: **Combining bayes classification and point group symmetry under boolean framework for enhanced protein quaternary structure inference.** *Structure London England 1993* 2011, **19:**304–12.
12. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state.** *J Mol Biol* 2007, **372:**774–97.
13. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL: **Statistical analysis of interface similarity in crystals of homologous proteins.** *J Mol Biol* 2008, **381:**487–507.
14. Xu Q, Dunbrack RL: **The protein common interface database (ProtCID)–a comprehensive database of interactions of homologous proteins in multiple crystal forms.** *Nuc Acids Res* 2011, **39:**D761–70.
15. Schärer MA, Grütter MG, Capitani G: **CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts.** *Proteins* 2010, **78:**2707–2713.
16. Hurst LD: **The Ka/Ks ratio: diagnosing the form of sequence evolution.** *Trends Genet* 2002, **18:**486.
17. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42:**108–124.
18. Ponstingl H, Kabir T, Thornton JM: **Automatic inference of protein quaternary structure from crystals.** *J Appl Crystallogr* 2003, **36:**1116–1122.
19. Brünger AT: **Free R value: cross-validation in crystallography.** *Methods Enzymol* 1997, **277:**366–96.
20. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci USA* 1996, **93:**13.
21. Ponstingl H, Henrick K, Thornton JM: **Discriminating between homodimeric and monomeric proteins in the crystalline state.** *Proteins* 2000, **41:**47–57.
22. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces.** *J Mol Biol* 2004, **336:**943–55.
23. Levy ED: **PiQSi: protein quaternary structure investigation.** *Structure* 2007, **15:**1364–7.
24. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites.** *J Mol Biol* 1999, **285:**2177–98.
25. Chakrabarti P, Janin J: **Dissecting protein–protein recognition sites.** *Proteins* 2002, **47:**334–343.
26. Levy ED: **A simple definition of structural regions in proteins and its use in analyzing interface evolution.** *J Mol Biol* 2010, **403:**660–670.
27. Bordner AJ, Gorin A: **Comprehensive inventory of protein complexes in the protein data bank from consistent classification of interfaces.** *BMC Bioinforma* 2008, **9:**234.
28. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* 1998, **280:**1–9.
29. Kryshtafovych A, Fidelis K, Moult J: **CASP9 results compared to those of previous casp experiments.** *Proteins* 2011, **79**(Suppl 1):196–207.
30. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science* 2001, **294:**93–96.
31. Poupon A, Janin J: **Analysis and prediction of protein quaternary structure.** *In Molecular Biology. Clifton, NJ: Humana Press;* 2010, **609:**349–364.
32. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257:**342–58.
33. Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N: **The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures.** *Proteins* 2005, **58:**610–617.
34. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9:**56–68.
35. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13:**190–202.
36. Krissinel E: *Detection of protein assemblies in crystals*. Life Sciences: Computational; 2005.
37. Kato HE, Zhang F, Yizhar O, Ramakrishnan C, Nishizawa T, Hirata K, Ito J, Aita Y, Tsukazaki T, Hayashi S, Hegemann P, Maturana AD, Ishitani R, Deisseroth K, Nureki O: **Crystal structure of the channelrhodopsin light-gated cation channel.** *Nature* 2012, **482:**369–74.
38. Liu S, Li Q, Lai L: **A combinatorial score to distinguish biological and nonbiological protein–protein interfaces.** *Proteins* 2006, **64:**68–78.
39. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biol* 2011, **12:**125.
40. **Jmol: an open-source java viewer for chemical structures in 3D**. http://www.jmol.org/.
41. DeLano WL: **The PyMOL Molecular Graphics System,** *Version 1.5 Schrödinger, LLC*. 2002 (see www.pymol.org).

42. Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J: **An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor.** *Cell* 2006, **125:**1137–49.
43. Stamos J, Sliwkowski MX, Eigenbrot C: **Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor.** *J Biol Chem* 2002, **277:**46265–72.
44. Landau M, Fleishman SJ, Ben-Tal N: **A putative mechanism for downregulation of the catalytic activity of the EGF receptor via direct contact between its kinase and C-terminal domains.** *Structure* 2004, **12:**2265–75.
45. Monod J, Wyman J, Changeux J-P: **On the nature of allosteric transitions: A plausible model.** *J Mol Biol* 1965, **12:**88–118.
46. Eliot AC, Kirsch JF: **Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations.** *Annu Rev Biochem* 2004, **73:**383–415.
47. Shrake A, Rupley JA: **Environment and exposure to solvent of protein atoms.** *Lysozyme and insulin. J Mol Biol* 1973, **79:**351–71.
48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nuc Acids Res* 1997, **25:**3389–3402.
49. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.
50. Murphy LR, Wallqvist A, Levy RM: **Simplified amino acid alphabets for protein fold recognition and implications for folding.** *Protein Eng* 2000, **13:**149–52.
51. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **Dissecting subunit interfaces in homodimeric proteins.** *Proteins* 2003, **53:**708–19.
52. Team RDC: **R: a language and environment for statistical computing**. Vienna Austria: R Foundation for Statistical Computing; 2010.

doi:10.1186/1471-2105-13-334
Cite this article as: Duarte *et al.*: Protein interface classification by evolutionary analysis. *BMC Bioinformatics* 2012 **13**:334.

Supplementary tables

Supplementary Table 1: DCxtal dataset

entry	area	resolution	xtal interfaces used	evidence	reference	comments
2q7d	1783.63	1.6	1	SEC, SLS, comparison to 2odt	17616525	Monomer. The paper states exactly that "biophysical methods like gel filtration and static light scattering indicate a monomeric state" but data is not shown. The same protein was solved by SGC 2odt and does not have that interface at all. Note that the HIS tag from both chains is involved in interface, however only in part of the rim.
2gas	1566.9	1.6	1	SEC	16600295	Paper clearly states it is a monomer, they used SEC and show data.
3c8y	1522.73	1.4	1		2173950, 2544883	Iron hydrogenase 1 from *Clostridium pasteurianum*. The review says it is a monomer, quoting this paper, but it's not clear what's the evidence there. Anyway all publications after that seem to assume that it is surely a monomer, e.g. 9836629 which is the primary reference for 1feh, same protein as this one solved before . That one contains also the same interface (but it's also the same crystal form).
3mhz	1516.29	1.7	1	SEC	20672855,9039918,15243628, 17335404, 4966833	The paper and the (extensive) literature on this PA anthrax protein assume that it is a well-known fact that it is a monomer (82KDa). It's cleaved and the 62KDa part heptamerizes to form a pore. First gel filtration in 4966833, established molecular weight of ~ 100 KDa, thus monomer. Entry 1t6b is the same protein complexed with a human cell receptor and also contains a large putative crystal interface.
2wbq	1515.87	1.1	1	SEC	19490124, 15368580	No info in paper (though they do mention they did gel filtration), but in a previous paper (15368580) characterized as monomer by SEC. Homologous to another protein in this list 2og5 (30% id with good structural conservation). The interface in both cases is the same one anyway (even though space groups are different).
3aap	1382.3	1.6	1	SEC	20159467	Monomer. Authors say so in paper, they used SEC but data not shown

2yz1	1378.66	1.4	1	SEC, AUC	18045614	This ligand binding domain (extracellular domain of a membrane protein) seems to be a monomer in solution as shown by SEC (no data shown) and AUC (sedimentation equilibrium, data shown). As further hint the authors note that the interface is antiparallel and thus unlikely to be formed by a transmembrane protein on a single cell surface. The full length including transmembrane domain could be a dimer.
2ipi	1329.37	1.7	1,2	SEC	17395717	Monomer following publication, they did SEC, no data shown. Four chains in the a.u., with two nearly identical large interfaces (B+A)
3cj1	1124.4	1.7	1	native PAGE, SEC, DLS	17910474	Brenda assigns as dimer. 17910474 claims monomer by native PAGE (data shown) SEC and dynamic light scattering (data not shown)
2eyi	1111.83	1.7	1	DLS, SEC	12657793	a-actinin binding domain, monomer, dimerization domain before, scheme
1pp3	1110.48	1.6	1	SEC	DOI: 10.1021/cg800616q	Monomer by SEC
1ynq	1101.76	1.3	1		16242712	Personal communication by XD Li
2w20	1100.67	1.5	1	SEC	18765901	Truncated version of protein, predominantly monomer, full length may be dimer.
3mg1	1099.32	1.6	1	SEC	20368334	Authors say it is a monomer based on SEC, they show data and looks like well calibrated. Homolog (~70% seq id) 1m98 has the same interface and the paper says that it is a dimer based on SEC, but they don't show the data. Thus we call monomer following the first paper as they seem to have good data.
3cu9	1098.4	1.2	1	SEC	19505290	Monomer by SEC, so the authors claim in the paper, without showing data. Originally the entry in the list was the mutant (from the same study) 3d5y but we replaced it by the wild type 3cu9. Both contain the same interface
1n45	1097.6	1.5	1	CCL, FRET	19556236	Forms dimers/oligomers in ER, monomer when it does not have TM helix.
2hlq	1087.36	1.5	1	SEC	17094948 16982201	Seems monomeric following their very detailed paper (16982201) on purification of this protein. They did SEC and show data (construct D 32-131 is the one used for crystallisation).
1s83	1063.55	1.3	1	SEC	19585993	Monomeric, bovine pancreatic trypsin was used as a standard in SDS-PAGE.
3go5	1063.14	1.4	1	SEC	20399190	Monomer claim the authors, by SEC but no data shown.
3fwk	1046.84	1.2	1	SEC	19375431	Authors claim monomer by SEC. They show the data in supplementary material (Fig. S1). Homolog 2wsi (~50% seq id and with very good structural conservation) does not have that interface.
1gpi	1039.54	1.3	1		11743726	Circumstantial evidence: crystal structure is truncated enzyme, additional residues for full length would disrupt interaction.

1lxk	1033.37	1.5	1	SEC	9365	Monomer in Bacillus sp.
2j0p	1017.8	1.7	1	SEC	16943192	Authors claim monomer by SEC. They present the data (Fig. S2 of supplementary) and they really did a very good job: well calibrated, done at different protein and salt concentrations.
1xgk	1017.46	1.4	1	SEC, DSC unfolding	11679757 15537757	Seems to be a monomer. Determined by SEC (11679757), some data given but no plots. Also main publication for this structure 15537757 claims monomer through DSC unfolding. (Note that the 2 names used for the species of this NmrA protein: Emericella nidulans and Aspergillus nidulans refer to the same fungal species, see wikipedia)
3gkj	1014.66	1.6	1	AUC (SV)	19563754	Authors claim monomer by AUC. Previously supposed to be dimer (17989072) by SEC, but authors say that "Previous results with glycosylated proteins have shown anomalous migration behavior on size exclusion chromatography", that's why they did the AUC.
3m66	1009.59	1.6	1		9118945	MTERF3, binds as monomer to DNA, in 16787637 it is stated that it is a monomeric protein
2wsa	1007.3	1.6	1		20036251	Authors state that it is monomeric. Brenda has orthologs characterized by SEC all monomeric.
3hzl	1005.4	1.6	1		17697998	Stated by the authors to exist as monomer in solution.
1w9q	1005.13	1.7	1	NMR	16533050 15698575	Contains not interesting peptidic ligand (chain S). The second paper 15698575 is a full NMR study of the protein's quaternary structure.
3h30	1328.83	1.6	1,2,3		11574463, 17084631	Comparison with 1JWH shows that the dimerization of the CK2 reg. subunit is NOT that observed in 3h30 (heterotetramer, as confirmed by BRENDA). Thus, accepted
2wbf	1323.25	1.6	1	SEC	13679369, 19591843	Fragment of domain SERA5PE of protein SERA5 of Plasmodium Falciparum. Domain SERA5PE is monomer by SEC (13679369)
2j46	1235.11	1.1	1,2	SEC	doi:10.1016/S0167-4838(02)00287-X	Elutes as monomer
3kk8	1233.14	1.7	1	MALS	20139983	Full length forms tetradecamer, kinase domain is monomer (MALS)
1ejd	1227.05	1.6	1	native PAGE	1577165	Characterized as monomer
3gvo	1222.17	1.6	1	SEC,DLS (for homologue, 79%)	19372537 11303521	Mouse Pumilio-2 Puf Domain. High sequence identity with human (~90%) and drosophila (79%). For Drosophila (3h3d) shown to be monomer by SEC and DLS (11303521). The Drosophila structure (3h3d) is a different crystal form and doesn't have the interface. Human homologues are same crystal form as 3gvo with same interface (e.g. 1m8z).

1j96	1218.8	1.3	1	SEC (homologue, 68%)	11514561	Not clear, according to BRENDA in *Rattus norvegicus* monomer by gel filtration (6435601). Seq id is 68%. Primary citation claims monomer in solution (data not shown). Keep
2xov	1199.74	1.7	1		21256137	Transmembrane protease, up and down arrangement, clearly a crystal contact
1so7	1189.06	1.5	1	SEC, disc gel electrophoresis	6735353	Monomer by gel filtration and disc electrophoresis
2cki	1185.89	1.7	1	SEC	16627477	Monomer, assessed by gel filtration
1lqt	1180.32	1.1	1	SEC	12071965	Monomer, assessed by gel filtration
2ow9	1178.46	1.7	1	SEC	17623656, 9790892	Literature found in MEROPS. Same sequence with similar boundaries analyzed by gel-filtration: monomer
2f37	1158.58	1.7	1	SEC	16882997	Ankyrin repeat, monomer in solution, no data shown
2z6o	1158.39	1.6	1	NMR, SLS	19101823	Ubiquitin protein ligase, monomeric as stated in 19101823 by SLS (data not shown) and rotational tumbling correlation time in NMR.
1ueb	1150.1	1.7	1	AUC, LS	15210970	Monomer by AUC and light scattering
3b37	1148.82	1.7	1	SEC	ISSN: 0002-1369 (AGRICULTURAL AND BIOLOGICAL CHEMISTRY, 1988, 52:217)	Monomer by SEC
1wly	1143.41	1.3	1	SEC	15781461	Monomer by SEC (stated in 15781461). However E. coli homolog 1qor (~40%id and very well conserved structurally) has exactly the same interface and they claim it is a dimer in the paper without presenting much experimental evidence.
1lf2	2171.42	1.8	1	SEC, AUC	12454457 17040901	EC 3.4.23.39. Interesting case: dimer in several crystal forms, monomer in solution under the conditions where protein is active, but higher oligomers can form irreversibly. Keep.
2qb5	1790.2	1.8	1	SEC, SLS, comparison to 2odt	17616525	Authors claim monomer in paper from SEC and SLS but data not shown. There's a HIS tag in the N terminal which is involved slightly in the interface. But most of the interface is formed by the C-terminal. Comparing to same structure but different construct 2odt from SGC, that one does not have this interface at all.
1zlq	1678.52	1.8	1	SEC, AUC	7867647 12960164	Was characterized as monomer in earlier publication 7867647 by SEC, data not shown. Then by AUC in this other paper 12960164 (reference of structure 1uiu). Also comparing to same protein 1uiu (first structure solved for this NikA protein), the interface is not present in that one.

2yvw	1522.84	1.8	1	SEC	1577165	SG. There's not much info about this Aquifex aeolicus MurA enzyme, but there are a few very well structurally conserved (~40% seq id, ~1A rmsd) homologs from E. coli, E. cloacae and others. They all seem to be monomers following Brenda. This early paper 1577165 shows evidence for E. cloacae (3kqa) being a monomer by SEC. The various homologs have different crystal forms and none of the others have this interface. Thus monomer.
1d3h	1483.64	1.8	1	SEC	10673429	Authors state in paper that it is a monomer by analytical gel filtration chromatography. This is human DHODH enzyme, the structurally closely related homolog from Lactococus Lactis (1jue) is a homodimer, the dimerization interface of that one is different from this one
2e1v	1653.41	1.8	1	SEC	17383962	Authors state that it is monomeric by gel filtration.
3n5c	1457.8	1.8	1	SAXS, SEC-MALS	20709080	The authors do analysis by SAXS and by SEC-MALS presenting data and a lot of evidence. This seems to be a very clear monomer.
2wbm	1439.53	1.8	1	AUC	19454024	No data shown for sedimentation equilibrium experiment. Authors state "no dimer has been observed for any of the available structures of afSBDS, which show different crystal packings."
3c1d	1427.07	1.8	1	AUC, SEC, CL	18650935	Sedimentation velocity (data shown), homo-bifunctional chemical crosslinking (not shown)
1ndb	1386.57	1.8	1	SEC	12526798	SEC, data not shown. Consistent with publication by Ramsay
2x26	1377.16	1.8	1	SEC	20383006	Interesting interaction: as in fim proteins there is a 13-amino-acid tail (but C-terminal) that folds within the cleft of the next monomer (cloning artifact: coupling of protein to GFP for estimation of expression levels). Protein runs as monomer in SEC (data not shown). No homologs in PDB
3kh7	1333.8	1.8	1		20544959	Most likely monomer, though no data provided. Homolog from E.coli is monomer 11843181, rmsd between 3kh7 (pa) and 2b1k (ec) 0.95 A
1toa	1294.51	1.8	1	SEC	10404217, 10400603	Monomer by SEC
1ffr	1226.41	1.8	1	SEC	Annals of Microbiology, 55 (3) 213-218 (2005)	Chitinase from S. Marcenses (bacterium), uniprot P07254. It's identical in sequence to Sanguibacter C4's chitinase (uniprot Q2V9S9) and that one is monomer by SEC with data (Tao YONG, Jin HONG, Long ZHANGFU, Zhang LI, Ding XIUQIONG, Tao KE, Ge SHAORONG, Liu SHIGUI, Annals of Microbiology, 55 (3) 213-218 (2005), Purification and characterization of an extracellular chitinase produced by bacterium C4) (link to pdf). Brenda also has many monomers for bacterial chitinases
1vqq	1215.7	1.8	1	SEC	8163510	Seems to be monomeric by SEC

1n4g	1162.15	1.8	1	MALLS	20621636	CYP121, it's monomer
3ita	1142.07	1.8	1	native PAGE	19807181	Publication claims that PBP6 behaves as monomer in solution
1woq	1129.82	1.8	1	SEC	12839753	15377666 publication claims monomer, evidence given in 12839753
1cqx	1110.84	1.8	1	SEC	8557026, 218634	Authors claim structure to be monomeric but no evidence given. See 218634 for clear SEC evidence.
2eqa	1105.25	1.8	1	LS	18004774	As described in publication
3mhj	1102.84	1.8	1	SEC	18436240	Human tankyrase 2, most likely monomer by comparison with tankyrase 1 2RF5 (evidence for tankyrase 1 given in 18436240), sequence identity is about 72%
2fgz	1091.03	1.8	1	Homology	16650854	Crystal Structure Analysis of apo pullulanase from Klebsiella pneumoniae, monomer in Klebsiella aerogenes (seq id 85%)
1fpo	1087.78	1.8	1	SEC	9144776	Protein seems larger than monomer, but smaller than dimer. interpreted as non-globular by authors. It makes sense if compared to structure
3lvd	1082.92	1.8	1	SEC	20220148 12693991	Green fluorescent protein mutant (aceGFP-G222E), wt is monomer by SEC; since G222E does not affect interface, accept
3irb	1060.05	1.8	1	SLS, SEC	20944206	PISA seems to predict 2mer or 4mer, experiments seem to tell otherwise, no data shown
1utj	1057.15	1.8	1	SEC	8896331	Trypsin of Salmo, SEC mentioned in publication
3f0o	1043.2	1.8	1	SEC	3542021	Monomer in solution, but crystal interface looks quite real
2h44	1038.64	1.8	1	SEC	16735511	Monomeric fragment (535-860). SEC was done on 508-865, the full-length protein is dimeric
1t8g	1029.16	1.8	1	SEC	15340171	Phage T4 lysozyme mutant L32A/L33A/T34A/C54T/C97A/E108V. Lysozyme is a monomer and the author state that "soluble and monomeric as judged by elusion profiles from sizing columns (data not shown)"
1g6a	1019.59	1.8	1	SEC	11148033, 235307	PSE-4 beta-lactamase, monomer by SEC (the authors quote an early paper where there is a calibrated SEC)
2zyr	1009.28	1.8	1	SEC, active site titration with [3H]DFP, native PAGE	19447113,, 10620337	AFL from *Archeoglobus fulgidus*. Crystallizes also in a form with only one monomer per a.u. (2zys), shares interface with 2zyr. Monomeric as determined by previous biochemical study
3els	1005.04	1.8	1	SEC	19010333	Clear chromatogram with calibration, monomer

Supplementary Table 2: DCbio dataset

entry	area	resolution	assembly	bio interfaces	evidence	reference	comments
2fwv	2085.26	1.7	2	1	DLS, SEC	17172346	It's a dimer, wrongly annotated as monomer. They check it with several methods in the paper: DLS, SEC (showing data for both) and homology. PISA predicts wrongly a tetramer as the most likely assembly
1ytq	1950.11	1.7	2	1	SEC, DLS	17327390	It's a dimer, wrongly annotated as monomer. The abstract says it already and they show it in the paper with SEC and dynamic light scattering.
1v2x	1705.29	1.5	2	1	AUC	15062082	It's a dimer, wrongly annotated as monomer. In paper they state they did AUC but don't show the data. Strangely the BSA they quote in the paper is 1353 A2 instead of our value (or PISA's) ~1700 A2.
1pkh	1673.65	1.4	3	1,2	SEC	12909016	It's a hexamer, wrongly annotated as monomer. Clear from crystal and also proofed with SEC in paper. Actually it's a dimer of trimers. They proof in the paper that it's a hexamer and not a trimer with SEC, but it's not totally clear whether the hexamer is real (could be simply low affinity). As the 100% clear one is the trimer (interfaces 1,2) we'll take those 2 as bio, interface 4 would be the one corresponding to the dimer of trimers oligomerisation.
1r5y	1554.08	1.2	2	1	SEC, non-covalent MS	7665516 19627989	Zymomonas mobilis (this one) was considered a monomer (SEC, data not shown), but a recent paper shows that it is a dimer by non-covalent MS and that it is the dimer that binds one substrate tRNA molecule at a time.
1kq3	1486.08	1.5	4	1,2	SEC,EM,homology	11566129 11134946	A tetramer. No publication (SG). But homolog 1jq5 is a tetramer (clearly stated in paper with support from EM and SEC data). Identity is ~40% but structure is very well conserved. They both contain the same interface and furthermore they both crystallize in same space group (I 4 2 2). Actually there's a tetramer/octamer discussion for 1jq5, in 11566129 they say an octmer fits the EM data but looks not so convincing, in 11134946 theres SEC evidence for octamer, but they also mention an earlier paper which claims tetramer. Taking only interfaces 1,2 (tetramer) and not 3 (the octamer one)
1jq5	1318.52	1.7	4	1,2	SEC,EM	11566129 11134946	See 1kq3 above.
2h7i	1481.31	1.6	4	1,2	SEC	17588773	Tetramer. The protein has been solved many times (26 entries with 100% id in PDB). The paper from 2idz explicitly says it's a homotetramer from SEC data (not showing data).
2nzl	1299.14	1.4	4	1	SEC	17669354	A tetramer. SEC evidence in this paper found in reference of main paper of entry 2W0U (same protein as this one).
1uj6	1281.61	1.7	2	1	SEC,DLC	13679361	Homodimer by SEC and DLC according to main reference of structure
1ju3	1268.31	1.6	2	1	SEC	11742345 20436035	Dimer by SEC (20436035). Main reference and PDB annotation say monomer
1sml	1258.8	1.7	4	1,2	SEC,AUC	9811546	Wrong annotation: tetramer by SEC (older references from main paper) and AUC sedimentation equilibrium with data shown (9811546)
3fah	1240.29	1.7	2	1	AUC,SEC	8354279	Wrong annotation, characterized as dimer.
1lzl	1207.56	1.3	2	1	SEC	12421810	Dimer by SEC
3iue	1150.47	1.7	2	1	SEC	11669627	Dimer by SEC in 11669627

1vk5	1144.85	1.6	2	1	SEC	16511118	"Gel filtration indicates homodimer" (data not shown)
2exb	1899.64	1.8	2	1	AUC	16411754	Dimer, the paper says that "it forms a tightly bound dimer" from AUC analysis (sedimentation and equilibrium), but data not shown.
3a2q	1745.79	1.8	2	1	SEC,AUC	923591	It's a homodimer by SEC and AUC, see 923591
3h6d	1665.4	1.8	3	1			It's a homotrimer. Couldn't find the citation with the exact experimental evidence, but it is assumed by all authors that it is known to be a trimer. There are 10 structures for this protein in the PDB. All showing the trimer, e.g. 1snf which is also a different crystal form
1s2z	1621.4	1.8	2	1	AUC	2835096	It's a homodimer by AUC. Has 2 large (>1500A2) interfaces. Both are in all 10 structures for this protein in PDB, all of them same crystal form though (see protcid)
2z1n	1586.77	1.8	2	1	SEC	18175326	Dimer or tetramer by gel filtration (data not shown). Actually there are 2 different interfaces that form 2 possible dimers and altogether a possible tetramer. The gel filtration gives a ratio 10:1 dimer to tetramer. But it's not known which one of the 2 interfaces is the possible dimer observed. We take only interface 1 as a sure bio to stay in the safe side
2vef	1556.35	1.8	2	1	SEC	3114239	Dimer by gel filtration, see 3114239. Other homologs from similar organisms (e.g. 1eye from M.Tuberculosis) superpose well, have same interface and are known to be dimers, see 11007651
1bs1	1483.09	1.8	2	1	SEC	4921568	Was earlier characterized as homodimer. It's been solved many times and all the structures in the PDB have the same interface
3lw6	1010.95	1.8	2	1	SEC,UDP binding stoichiometry	20236943 19032152	Catalytic domain of Drosophila beta1,4-galactosyltransferase-7 (73% similarity to human). For human enzyme (delta 1-81) gel filtration and UDP binding stoichiometry show it is a dimer. Interestingly this entry (Drosophila) contains quite a long N-terminal expression tag taken from the homologous bovine enzyme. Anyway the tag is not even seen in the density and in any case doesn't come close to the interface.
3f3e	1270.99	1.8	2	1	analogy to serotonin receptor	16041361	Dimer according to 16041361
2vr4	1265.02	1.8	2	1	SEC,DLS	17287210	Dimer by SEC and DLC data not shown, reference is of same protein solved earlier by same group (2je8). Brenda has dimer or higher oligomer for same EC (3.2.1.25) in related organisms
1x7v	1191.21	1.8	2	1	SEC	16049913	"The PA3566 protein crystallizes as a trimer, although the functional unit is most likely a dimer, as are other members of this structural superfamily." confirmed by SEC
1f2d	1892.91	2.0	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
1lw4	1857.20	1.9	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
1n8p	1969.27	2.6	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
1qop	1531.10	1.4	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
2aq6	1195.11	1.7	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch

2bhs	1487.99	2.7	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
2cft	944.77	1.8	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
2e7j	1668.16	2.4	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
2ecq	1623.08	1.9	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
2rkb	1104.54	2.8	2	1		15189147	PLP enzyme, known to be dimer or higher oligomer, see review 15189147 by Eliot and Kirsch
1eej	856.37	1.90	2	1	SEC	7536035	Well characterized as homodimer, see 7536035.
1o17	933.97	2.05	2	1,2	SEC	11298741	Well characterized as homodimer, see 11298741. Two copies of the interface in the ASU.
1ze3	1473.17 1093.65	1.84	3	1,2	SEC	15920478	Ternary complex of FimD-C-H. Characterized as complex by SEC
3d36	665	2.03	4	1,3	Mutagenesis	19101565	Well characterized complex (on interface 3) of KinB (chains A,B) and small protein Sda (chain C). Interface 1 is for homodimer A+B. Interface 2 is really interesting because it is a crystal contact that replaces the bio contact of interface 3 in the other (symmetric) side of the molecule. The paper is very thorough and they check the interface with different techniques, including mutagenesis.
3r0n	901.84	1.3	2	1	SEC,Native PAGE	22547693	Immunoglobulin variable domain of Nectin-2. The reference is the primary citation of same protein 4dfh
3da8	911.22	1.3	2	1	DLS,SEC	19394344	Authors did both DLS and SEC but show no data, they solved 2 structures for the same protein in different crystal forms, both have the interface providing further evidence.
2c4w	914.5	1.6	12	1,2	AUC	1554351	Type II DHQase from H. Pylori. It's a dodecamer by similarity to 2y71 (M. Tuberculosis, only 30% seq id but amazing structural conservation at the dodecamer level, pymol aligns the 2 dodecamers downloaded from PISA with 1.4 rmsd) and to 1gu0 (35%id, very good structural conservation at dodecameric level). 2y71 and 1gu0 have been well characterized biophysically (1554351) as dodecamers. Interfaces 1 and 2 are engaged to form the dodecamer.
3jrz	933.25	1.7	2	1	NMR	19959472	They solved both the crystal and the NMR. In the abstract they say that it is a dimer in solution, but the only proof of it in the paper is from the NMR NOE peaks that are assigned to inter/intra molecular ones via the X-ray structure. Additional clues: 1) 2 crystal forms solved have the same interface, 2) a well conserved homolog of E coli 3hpw (~40%id) also has the same interface
3f6q	949.99	1.6	2	1	SEC	19074270	Citing the paper: "the complex remained intact through further rounds of ion-exchange and size-exclusion chromatography". They also tried mutagenesis on several residues, especially mutant F42A caused near complete loss of binding. By the way, the F42 is the only core residue in that side of the interface
2wxd	960.02	1.6	2	1	SEC,AUC	2369130 9195886	Well charaterized dimer, original reference is 2369130. Solved a few times in the PDB.
2bz6	981	1.6				16621574	FACTOR VIIA heavy and light chain linked by S-S bridge. The original single chain is cleaved by other factors of the extrinsic blood coagulation pathway into the heavy and light chain (P70375 UniProt)

2d0d	1003.06	1.7	2	1	SEC	16233251	Homodimer by SEC. This is a mutant but I couldn't find the WT, there are a few other mutants in PDB. In any case the mutation is not in interface.
3cm3	1012.27	1.3	2	1	AUC-SV	19211553	Homodimer by AUC - sedimentation velocity, they show data.
3o1n	1017.98	1	2	1	SEC	21291284 8216229	AroD from *Salmonella enterica serovar typhimurium* ; AroD from *Salmonella typhi*, 100% id, is dimer by SEC (8216229)
3h0n	1018.4	1.5	2	1	SEC,SLS	20944211	SG, new fold (ABATE domain), published.
2v52	1030.07	1.5	2	1	fluorescence anisotropy	19008859	MAL-RPEL2, Kd by fluorescence anisotropy 0.1 uM.
1zlh	1033.56	1.7	2	1	inhibition kinetics, titration	15961103 15561703	Heterocomplex carboxypeptidase A1:proteinaceous inhibitor. It binds 1:1 (titration) with Ki ~ 1nM
2dvn	1045.17	1.6	2	1	DLS	18062990	
3ovp	1060.91	1.7	2	1	SEC, AUC	20923965	
2i7d	1077.94	1.2	2	1	SEC	2157703	The main reference for this entry is 17985935
3bzl	1085.28	1.2	2	1	Dimer because the two chains originate from a self-cleaved protein which retains its fold	18451864	Special case: this C-terminal domain of the EscU protein from the type III secretion system of Gram-negative bacteria self-cleaves into two polypeptides and the two chains stay together. Originally the entry we had from the filtering was 3bzy, but it is a mutant, wt is 3BZL (1.71 A).
2car	1099.46	1.1	2	1	SEC	17138556 11278832	Clear dimer by SEC, they show the calibration curve
3gus	1221.37	1.5	2	1		16597834 16399376	Very well studied human glutathione S-transferase enzyme (2.5.1.18) type p1, standard name hGSTP1-1. Known to be homodimeric, could not find the exact reference. Many structures in PDB, all having the interface.
3jyo	1224.52	1	2	1	SEC	18566515	Homodimer by SEC
1uz3	1228.26	1.1	2	1	SEC,AUC	15978617	N-terminal domain of EMSY protein (Q7Z589), homodimer in solution by SEC, AUC and even yeast two-hybrid evidence. Kd established at ~2uM
2wtm	1243.68	1.6	2	1	DLS	20058325	Homodimer in solution by DLS. They solved 2 different crystal forms of the protein, both containing the interface.
2a5l	1252.84	1.7	2	1	AUC,SEC	16322580 9694845	Dimer-tetramer equilibrium, we take thus the 1st interface as valid (corresponding hopefully to dimer). Another structure within 95% id has been solved: 1zwl. Both have in common interfaces 1 and 2 and are 2 different crystal forms.
2y39	1294.51	1.4	2	1	SEC,AUC,NMR	18825506	A homodimer by SEC, AUC and NMR. 3epv is the same protein with same interface except that it is slightly bigger, maybe because of different bound metal? In any case the evidence looks quite solid.
2ab0	1296.42	1.1	2	1	DLS	16181642	Protein YajL from E coli, dimer by DLS and homology to structurally very close human protein DJ-1 (1p5f)
3epw	1302.63	1.3	2	1	SEC	11292348	Homodimer by SEC. Solved a few times in PDB. Reference is from first one 1hoz, which lacks part of a helix in the interface (missing density I guess), resulting in a much smaller area for the interface.
2g2u	1304.7	1.6	2	1		21294157 16809340	Well studied complex of SHV-1 + BLIP. Kds have been measured for WT (2uM) and mutants. We had originally the mutant 3n4i and I have replaced it by the WT 2g2u.

2vvt	1340.2	1.7	2	1	MALS,AUC	17568739	Homodimer in solution, from MALS and AUC. They did a thorough study on this and many related homologs from other bacteria, characterizing and crystallizing all of them. Seems very clear
1p5f	1343.76	1.1	2	1	SEC	12855764	Well characterized human DJ-1 protein. We have a homolog in this list (2ab0). They did SEC and show data. We originally had a mutant in the list (2rk3), I've replaced it by WT 1p5f at 1.1A resolution
3itf	1359.9	1.5	2	1	SEC,SGS	21239493	Homodimer by SEC and Sucrose Gradient Sedimentation, through them they determine a mass of 31KDa. All data shown
3kd2	1371.5	1.8	2	1	SEC,AUC-SV	20118260	Homodimer by SEC, AUC-SV, showing data, looks solid. Originally we had mutant 3kda in the list, replaced by WT 3kd2, resolution 1.8
2w6a	1394.15	1.4	2	1	MALLS, AUC-SE	19136011	Clear case
2y27	1395.2	1.6	2	1	SEC	21388965	Clear case

2.1 Appendix: algorithmic details

As an appendix to the *Duarte et al BMC Bioinformatics* (2012) publication presented in this thesis chapter, we will also elaborate in some detail about some of the algorithms used in the EPPIC software but not described in the paper's text. The algorithms described below are all implemented in Java in the OWL package, for which the source code is available in a Subversion repository at the URL svn://bioinformatics.org/svnroot/owl.

2.1.1 Interfaces enumeration

The problem here concerns finding out the unique protein-protein interfaces present in a given crystal lattice. The crystal lattice is given as an asymmetric unit containing one or more protein chains and a space group together with the unit cell parameters. From those it is possible to reconstruct the lattice and in principle it is straight forward to find the contacting subunits and then the unique set of them.

Each space group is described by a set of l operators, if we then have n distinct chains in the asymmetric unit that gives us a total of $m=nl$ subunits in the unit cell. In principle it suffices with checking the contacts from the first unit cell to all c=26 neighboring cells, though sometimes it is necessary to go as far as the second shell of neighboring cells. In any case an exhaustive search would require checking $n^2 + nmc$ inter-chain contacts. The algorithm for the actual inter-chain contact calculation is explained then in detail in the next section.

The two main shortcuts taken for the calculation are: a) using bounding boxes that are translated around before applying transformations to all coordinates, so that there is no need for translating the whole chain or performing inter-chain contact calculation when the chains are too far apart in space; b) using symmetry redundancy elimination: any two (rotation+translation) operators A, B for which $A \cdot B$ = *Identity* are equivalent and thus the second operator can be discarded after the first one has been trialed.

The interfaces enumeration algorithm thus proceeds as follows:

- First all pairwise possibilities within the asymmetric unit are tried, i.e. a total of $n(n-1)/2$ trials.
- Then contacts between the asymmetric unit chains and each of the other asymmetric unit symmetry-related copies in the unit cell are checked, using the symmetry redundancy elimination described. A maximum of nm trials are performed, fewer if symmetry redundancy exists for the space group.

- Finally contacts between the original asymmetric unit and all other asymmetric units of neighbouring unit cells are checked. In here we perform only translations as the rotations have already been calculated for the first unit cell. Before actually performing the translations, the bounding boxes alone are translated and checked to see if there is any overlap between the original AU and the others. If not the translation can be discarded directly. Symmetry redundancy is also checked in order not to repeat equivalent operators that already were tried. Thanks especially to the bounding boxes shortcuts it is possible to explore even the second cell neighbors, still staying within reasonable computing times.
- The list of interfaces found is then checked for duplicates, based on the sets of atoms and residues that were found to be in contact. A final list of unique interfaces is output.

The only remaining task is that of calculations of ASAs which is done with an implementation of the Shrake and Rupley algorithm [1] added to the OWL package. One needs to calculate the ASA of single chains ($ASA_{uncomplexed}$) and then that of the complexed pairs ($ASA_{complexed}$). So in theory it is enough with calculating the n single chains and the N pairs of chains (for each of the N interfaces found). In practice an issue with the sampling done in the Shrake and Rupley algorithm appears: the ASA of a chain will not coincide exactly with that of a rotation of it (with differences between 5-10%), which in the final BSA calculations ($ASA_{uncomplexed}$-$ASA_{complexed}$) can result in negative values for some surfaces. In order to avoid this, one needs to calculate ASAs per unique rotation (disregarding translations) and finally obtain the BSA values.

2.1.2 Inter-chain contacts calculation

The necessary step towards calculating an interface is to establish whether two protein chains are in contact. In this context "contact" is defined as any 2 atoms from both chains being within a pre-defined cutoff distance, which here we chose to be 6 Å.

In order to accomplish this, a novel algorithm implemented in the OWL library for calculation of inter-residue graphs (i.e. contact maps or intra-chain contacts calculation) was adapted to this purpose. In brief the algorithm uses a geometric hashing approach in order to avoid the calculation of all versus all atom distances (with complexity $O(n^2)$). The problem is eventually reduced to the calculation of all versus all distances within a grid cell plus neighboring ones. The steps are roughly:

- A 3-dimensional grid with cells of the size of the given contact cut-off is constructed.

- All atom coordinates are binned into the grid, by using a *floor* function which yields an integer 3-dimensional coordinate (corresponding to a grid cell) for each of the atoms.
- Then distances are calculated: 1) for all atoms within each cell 2) for atoms of one cell to atoms of each of the 3D neighboring cells.
- That results in a partial distance matrix that contains distances for any atom pair within the distance cut-off. From there the contact map follows.

The algorithm was then adapted to the inter-chain contact problem. Essentially the problem is the same except that in here the groups of atoms for which we want the distances come from two distinct sets *I* and *J* (from each of the two chains). The algorithm thus proceeds as follows:

- Before proceeding to grid the space one first calculates bounding boxes for each of the protein chains *I* and *J*. If the boxes do not overlap there is no contact and nothing left to be done.
- If the bounding boxes overlap then atoms need to be sorted into the grid analogously as it was done for the intra-chain contacts.
- For each grid cell distances between atoms of set *I* to atoms of set *J* need to be calculated. Then as before for every neighboring cell.
- We thus end up with a partial distance matrix containing distances within the given cut-off. From there the atom inter-chain graph follows.

3 An analysis of oligomerization interfaces in transmembrane proteins

An analysis of oligomerization interfaces in transmembrane proteins

Jose M. Duarte[1], Nikhil Biyani[1,2] and Guido Capitani[1*]

[1]Paul Scherrer Institut, CH-5232, Villigen PSI, Switzerland

[2]present address: Department of Chemistry, CH-8093, ETH Zurich, Switzerland

[*]Corresponding author:

Dr. Guido Capitani
Biomolecular Research
Paul Scherrer Institut
OFLC/110
5232 Villigen PSI, Switzerland
Phone: +41-56-3104722
Fax: +41-56-3105288
E-mail: guido.capitani@psi.ch

Running head: TMP interface analysis

Abstract

The amount of transmembrane protein structures solved to date is now large enough to attempt large scale analyses. We have compiled the first fully comprehensive set of validated transmembrane protein interfaces in order to study their features and assess what differentiates them from their soluble counterparts. The general features of TM interfaces do not differ much from those of soluble proteins: they are large, tightly packed and possess many interface core residues. In our set, membrane lipids were not found to significantly mediate protein-protein interfaces. Although no G protein-coupled receptor was included in the validated set, we analyzed the crystallographic dimerization interfaces proposed in the literature. We found that the putative dimer interfaces proposed for class A GPCRs do not show the usual patterns of stable biological interfaces, neither in terms of evolution nor of packing, thus they likely correspond to crystal interfaces. We cannot however rule out the possibility that they constitute transient or weak interfaces. In contrast we do observe a clear signature of biological interface for the proposed dimer of the class F human Smoothened receptor.

Keywords

protein structure; protein-protein interfaces; membrane proteins; eppic; lipids; GPCR

Introduction

Transmembrane proteins (TMPs) play a central role in biology. They are responsible for some of the most important functions of cells like signalling, transport and catalysis of important reactions. As a consequence, large efforts have been directed at the structural and functional analysis of TMPs. This feat required a series of technical and conceptual advances ranging from a detailed understanding of TMP reconstitution, purification and crystallization in detergents to approaches for optimization of data collection and radiation damage mitigation at synchrotron light sources.

Those efforts were highly successful and the number of available TMP structures in the Protein Data Bank kept increasing exponentially since the first structure determination in 1985 [1]. The last 15 years witnessed structure determination breakthroughs in TMP families that had previously resisted all efforts, like G-protein coupled receptors and ABC-transporters. According to Stephen White's MPSTRUC database of membrane proteins with known 3D structure (http://blanco.biomol.uci.edu/mpstruc), the number of unique membrane protein structures available as of 9th April 2013 is 393, a figure that includes not only TMPs but monotopic membrane proteins and some other membrane-associated proteins.

The abundance of high-quality structural data has made it possible to analyze membrane protein structures on a much larger scale and with a more solid foundation than only a few years ago. Recently studies have been performed on a variety of membrane protein-specific topics such as residue propensities at different membrane protein regions [2], lipid interactions [3], alpha-helical packing [4] or beta strand interactions [5].

This wealth of data makes it also possible to attempt a global analysis of protein-protein interactions and oligomerization in TMPs. To this end we compiled a manually curated dataset of membrane proteins for which the oligomeric state is well established from biophysical measurements and the structure has been determined at high resolution and quality. As analysis tool we used our Evolutionary Protein Protein Interface Classifier (EPPIC) [6], which we developed as a general approach to distinguish biological interfaces from lattice contacts in crystal structures. EPPIC depends on the availability of many homologues to the sequence of the protein being analyzed and its classification coverage and performance were retrospectively shown to improve, over a time span of 10 years, with the growth of the UniProt database. EPPIC reaches 90% accuracy on soluble proteins and we set out to assess its performance on our curated TMP dataset.

We also used our dataset to tackle an important issue in membrane protein structural biology: the presence and role of membrane lipids in TMP interfaces. The importance of lipids in membrane

protein folding and oligomerization has been subjected to study in the last years [7–9]. We would like to ascertain whether structural evidence exists that provides any insights into the role of lipids in the oligomerization of TM proteins.

Results and discussion

The dataset

We compiled a dataset of protein-protein interfaces that span the transmembrane region. In compiling such a dataset we adopted very strict selection criteria. First of all we restricted it to high resolution structures obtained from X-ray crystallography of 3-dimensional crystals in order to have a high quality and homogeneous dataset. The procedure required manual checking of the relevant literature to establish whether the oligomeric state of the TM proteins was known. Determining the oligomeric state of TM proteins experimentally is in itself a difficult task. Oligomerization can be measured in detergent via Size Exclusion Chromatography or Analytical Ultra Centrifugation as it would be the case for soluble proteins. However, the presence of detergent micelles and of the detergent belt around MPs complicates matters considerably. More sophisticated methods like FRET aim at determining the oligomerization state *in vivo* by using proteins tagged with chromophores and measuring the resonance energy transfer, very sensitive to distance [10, 11].

Owing to the filtering criteria several important cases were excluded from this dataset:

- Bacteriorhodopsin: bacteriorhodopsin and archaeal rhodopsins form membranes in vivo (purple membrane) which can be considered as natural 2D crystals [12]. Crystallographic studies find them associated as trimers in the native environment. However there is evidence of bacteriorhodopsin being a monomer in micelles [13] and even of it being functional in the monomeric state [14]. It was also solved via crystallization in bicelles [15] which resulted in a completely different crystal packing where no trimer association exists. Defining what constitutes an oligomer in the context of a 2D natural crystal thus becomes problematic. This precludes inclusion in the dataset since we need an independent non-crystallographic confirmation for the oligomerization state that it is not possible to provide for this case.
- GPCRs: there is a long-standing debate on GPCR oligomerization, see for instance [16–18]. Even though some experimental data are available and that some interfaces from crystal structures have been already proposed as possible dimerization interfaces [19–22] many questions remain open. Thus we decided not to include these interfaces in our dataset

of *bona fide* biologically relevant TM interfaces. We did, however, study in detail the different proposed dimer interfaces, as described in the GPCR section below.

- Mitochondrial ADP/ATP carrier: despite it being initially characterized as dimer it was later proven to be a monomer [23, 24] and thus the proposed lipid-mediated interface [25] was not included in this dataset. See also the Lipids and TM Interfaces section for further discussion.

The dataset comprises 62 oligomeric membrane protein structures with a total of 159 TM protein-protein interfaces, divided into the two subclasses: 46 from alpha class and 16 from beta class (see supplementary tables S1 and S3). This is, to our knowledge, the first fully comprehensive dataset of validated TM protein-protein interfaces from crystallography.

We must note that the oligomerization state of the proteins in the dataset was most of the times assessed in a detergent-solubilized state. We cannot rule out the possibility that in some cases solubilization with detergents alters the protein association occurring in the cell. In any case it remains very difficult with current technologies to reliably assess membrane protein oligomerization *in vivo*. Hence, this analysis represents a best effort providing a snapshot of the current knowledge.

Interface geometry and composition

The first analysis one can perform on the compiled dataset is in the geometry and composition of the interfaces. First of all we calculated the buried surfaces and number of interface core residues, which, as shown before for soluble proteins [6, 26] are a strong indication of an interface to be biological. Table S1 presents the data for all interfaces. Overall the geometry is quite similar to that of soluble proteins with large interfaces (only 7 interfaces below 900 $Å^2$) and many core residues (only 30 interfaces with 5 or fewer core residues, 15 with 4 or fewer). Following this, it seems clear that in terms of geometry (number of core residues) and packing the TM interfaces do not differ much from their soluble counterparts. To form stable complexes, protomers need to come together forming interpenetrating surfaces with many buried "hot-spots" residues. It thus seems that the tight-packing requirement is not only a consequence of the water environment but that it is also necessary in the context of the lipid bilayer.

We found only a few exceptions to the above observation, almost exclusively limited to light harvesting and photosynthetic complexes. Those two protein complexes represent special cases since they contain a very large amount of chlorophylls and carotenoids. Their oligomerization interfaces are not strictly protein-protein but rather protein-cofactor-protein ones.

Having confirmed that the packing of the TM interfaces is essentially like that of soluble ones, we studied whether any clear compositional differences in terms of the amino acid content can be observed. Figure 1 shows a comparison of amino acid frequencies at TM protein interfaces and of at soluble protein interfaces. The membrane proteins are sorted into their two major structural classes: alpha and beta. It is apparent that in terms of amino acid composition membrane and soluble interfaces are also quite similar, with the exception of alanine and glycine. Those two residues are clearly overrepresented in TM interfaces compared to soluble ones. Constraints imposed by helical packing are possible basis for this overrepresentation. It is known that in alpha helical TM domains small amino acids are important to enable helix packing [27]. Overrepresentation of Ala and Gly is less obviously connected to the subunit packing of beta TM proteins. We hypothesize that the flat interfaces formed by beta-to-beta packing also constrain the amino acids at the interface to be small as well as hydrophobic.

The data can also be presented in term of enrichments of the interface core residues versus the full protein (Figure 2) for both TM and soluble interfaces. The enrichments for most hydrophobic residues are clustered in the upper right quadrant while most charged or polar residues are clustered in the lower left quadrant. Thus for both soluble and TM interfaces the interface core residues are enriched in similar ways. Especially surprising is that no significant difference in enrichment can be seen for the hydrophobic residues in TM interfaces compared to soluble ones. This can be seen in a clearer way in Figure 3, where different properties of amino acids present at the interface cores are compared between the two groups of membrane and soluble proteins.

Lipids and TM interfaces

We then set out to determine whether membrane lipids can act as mediators in TM interfaces. We were not able to find any significant membrane lipid-mediated TM interface in the entire validated dataset. This is in agreement with what was found above in the packing. All interfaces present in the dataset are tightly packed, not leaving enough room for significant lipid interactions in the interfacial space.

The case of the electron transport megacomplexes deserves to be discussed in some detail. The cytochrome bc1, cytochrome c oxidase and Photosystems I and II are possibly the most complicated of the known TM protein structures in terms of subunit content, size, topology and lack of symmetric features. The interfaces present in these structures are in many cases not purely TM but spanning both the soluble and TM regions. Additionally, as is the case with light harvesting complexes, the presence of many porphyrin-based cofactors adds to the complexity. Some lipids are seen in the interfacial spaces, for instance in the cytochrome bc1 complex [PDB:

1ppj] a phosphatidylethanolamine molecule sits in a cavity where it interacts with chains C, D, E and J. However, the interaction of these chains occurs also through several extensive contacts on both intracellular and extracellular sides of the membrane.

Another interesting case is that of the bovine mitochondrial ADP/ATP carrier, where it was hypothesized that membrane lipids were essential for the interface formation. Initially it was characterized as a dimer [28]. Its first crystal structure [PDB: 1okc] [29] did not exhibit any plausible dimerization interfaces, since all of the crystal interfaces where either in an upside-down or head-to-tail orientation. Later on a new crystal structure was solved [PDB: 2c3e] where a very small interface (220 $Å^2$) mediated by cardiolipins was proposed as the dimerization interface, though the authors recognized that further experimental support was required [25]. The case was finally settled by Bamber et al, who demonstrated in two separate papers that the carrier is actually a monomer in detergent [24] and that it also functions as a monomer *in vivo* [23].

The case of bacteriorhodopsin, which we did not include in the dataset as discussed above, also deserves mentioning. A belt of lipids is seen in the high resolution crystal structures of bacteriorhodopsin from Lipidic Cubic Phase 3-dimensional crystals [PDB: 1m0k] [30], some of them located in the inter-trimer space. However the structure of a bacteriorhodopsin [PDB: 1kme] crystallized from bicelles [15] exhibits neither the trimeric arrangement nor the mediating lipids.

An important issue with membrane lipids is that of their high mobility and conformational flexibility, which makes it difficult to study them at atomic detail with crystallography. Indeed many of the crystallographic reported membrane lipids exhibit regions lacking electron density, which sometimes affects the interpretation and positioning of the entire ligand. In cases where chemically similar lipidic and detergent molecules are present in the crystal and ligand electron density is patchy it may even be challenging to distinguish a lipid from a detergent molecule. These issues belong to the broader problem of accurate electron density interpretation for non-protein ligands [31], which is often a challenge especially at the low resolution ranges typical of TM proteins. Independent validation for many ligands in the PDB has been performed and deposited in the Twilight server [31], where the ligand validity was objectively measured with a real space correlation coefficient (RSCC). Table S2 shows some prominent examples of Twilight RSCC values for lipids present in 11 representative alpha membrane proteins. Represented groups are bacteriorhodopsins, rhodopsins, potassium channel, ADP/ATP carrier, electron transport complexes, photosystems and light harvesting complexes. Out of 120 lipid molecules, 24 (20%) are below the Twilight threshold of RSCC 0.6, while 33% are below RSCC 0.7.

The above evidence speaks against lipids as mediators of biological contacts. However, they can be essential crystallization agents. It has been shown that for a membrane protein to be able to crystallize in a LCP mesophase, the lipidic composition of the cubic phase is key to obtain crystals [32]. Not only the hosting lipids that form the bulk of the mesophase are important but in some cases also adding "doping" lipids like cholesterol is necessary for a successful crystallization [33].

Classifying the interfaces with EPPIC

Once our dataset was compiled we used the method developed in our group [6] to attempt to computationally classify the TM interfaces as biologically relevant or not, as we previously did for soluble proteins. The EPPIC (Evolutionary Protein-Protein Interface Classifier) method relies on a combination of a simple geometrical indicator and of two evolutionary ones in order to classify an interface into biologically relevant or crystal lattice contact. It was demonstrated to work well on two validated sets of soluble proteins with an accuracy that is close to 90%.

Results for the TM dataset are presented in Table S1. The overall classification accuracy for this ensemble of *bona fide* biological interfaces is 80%, thus lower than that obtained earlier for soluble proteins [6]. It is worth mentioning that, in its current implementation, EPPIC analyzes interfaces in a pairwise manner only, without looking at the global assembly of interfaces present in the crystal and thus without taking the symmetry of the assembly into account. The symmetry of the assembly is indeed a very important factor, especially in membrane proteins where many of the known TM oligomers show highly symmetrical arrangements.

An example where the classification fails is in the structure of the rotor ring of Na-dependent F-ATP synthase [PDB: 2wgm]. The biological unit of this protein is a highly symmetric assembly with C11 point group symmetry, where chains consisting of a helical hairpin repeat 11 times around an axis. The core versus surface indicator cannot produce a prediction because of the few surface residues that are not interacting with other protomers. At the same time the rims of the interfaces happen to be very well conserved, possibly because some of the rim residues are involved in the sodium ion coordination. This results in high core versus rim values that fall out of the biological cut-off. The related structure of the rotor ring of a proton-dependent ATP synthase [PDB: 2wie] is misclassified by EPPIC in a very similar way, with analogous causes. The EPPIC method is known to have issues with small chains with little free surface like these cases. However the highly symmetric assembly of both cases would make a prediction based on symmetry considerations quite straightforward.

GPCR oligomerization

Oligomerization of G protein-coupled receptors is one of the most heavily debated topics related to TM interfaces [16, 34]. GPCRs constitute one of the largest protein families in animal genomes and are involved in receptor sensing and signal transduction processes, constituting one of the prime drug development targets with as much as 40% of drugs in the market targeting GPCRs. All members of the family share a very well conserved fold of 7 transmembrane helices and have evolved very fine selectivities in signal transduction. The family has been subdivided into 6 classes (class A to class F), being the class A of rhodopsin like receptors by far the most populated.

Most of the oligomerization debate has centered around the class A members where the evidence for oligomerization is least convincing. In contrast it is quite well established that class C receptors exist as stable dimers. Experimentally, FRET techniques have repeatedly been used for establishing association of receptors in the membrane. For instance evidence from FRET exists for some class A receptors, like the CXCR4 receptor which was shown to homodimerize or heterodimerize with the CCR2 receptor [35] [36].

Some dimer interfaces found by inspection of crystal structures have been proposed so far for several GPCRs. Distinguishing relevant interfaces in crystal structures is indeed a non-trivial task, which has been subject to a large amount of investigation [6, 37–40]. We decided to test the different proposed interfaces with the EPPIC method, which in principle is quite agnostic to crystallization artifacts, since it uses evolution to judge the biological relevance of an interface. The method is more powerful if abundant, relatively close sequence homologs are available for the alignments [6], especially if the distribution of identities in the homologs is uniform enough. Thus this makes the GPCR case a very suitable target for analysis with EPPIC, since sequence data are abundant for most family members. Predictions for this kind of case are *a priori* of a higher confidence.

We thus analyzed the different proposed interfaces, see Table 1:

- Bovine rhodopsin [PDB:2i35, 2i36, 2i37] [20]: two crystal forms were solved in the study, both containing a similar dimer interface. The trigonal crystal form has 3 molecules in the asymmetric unit and the dimer interface appears twice in that form, once between monomers A + B and another time between 2 symmetry-related C monomers. The buried surface area of the different dimers ranges from ~300 $Å^2$ to up to ~700 $Å^{2,}$ which is quite a significant variation, maybe attributable to the low resolution of the structures. In any case for all of them the packing in terms of number of core (fully buried) residues is typical for crystal contacts, ranging from 0 to 2 core residues counting both sides of the interface. The EPPIC evolutionary indicators, based on a large alignment of 105 homologs within 60%

identity, also suggest a crystal contact in all cases, even though in some of them poor packing does not allow the program to make a decision, as EPPIC requires at least 8 residues buried to 70% in order to produce a prediction.

It must be noted that the structures were determined at fairly low resolution: 3.7 Å, 4.1 Å and 4.2 Å, respectively. In that range of resolution it is quite difficult or impossible to properly model side chain rotamers, which may affect the packing quality of interfaces.

- Human CXCR4 chemokine receptor [PDB:3odu, 3oe0, 3oe6, 3oe8, 3oe9] [21]: five receptor structures, bound to a small-molecule antagonist or to a cyclic peptide, were solved in several crystal forms. The crystallization constructs were engineered for stability by insertion of a T4 lysozyme between TM helices V and VI. This way the lysozyme molecule becomes a soluble "domain" of the receptor. A dimerization interface can be seen in all of them in a parallel arrangement with poor packing (no core residues at all). The artificially inserted lysozyme "domain" is involved in some of those interfaces, which accounts for their larger size. We analyzed the evolutionary signal of the interfaces by stripping off the lysozyme from the atomic model and found a consistent crystal contact signature for all of them.

- Human κ-opioid receptor [PDB:4djh] [19]: the receptor was crystallized by engineering a T4 lysozyme fusion protein. An interface of 1000 $Å^2$, in which the lysozyme is not involved, was proposed as dimerization interface. In terms of packing the interface features the typical signature of crystal contacts with few core residues (only 2). Evolutionary analysis by EPPIC again yields a very clear crystal contact signal, based on an alignment of 102 homolog sequences within 60% identity of the human κ-opioid receptor.

- Turkey β1 adrenergic receptor [PDB:4gpo] [22]: in this case the crystallization strategy did not involve engineering of a fusion protein, but a set of stabilizing mutations plus removal of a loop. An interface of 800 $Å^2$ between NCS-related chains A and B was proposed to mediate receptor dimerization. Evolutionary analysis again indicates a clear crystal contact, based on an alignment of 49 homologs. Again it must be noted that the structure was solved at fairly low resolution.

In summary none of the proposed class A GPCR dimerization interfaces follow the patterns expected for high affinity biological TM interfaces in terms of geometrical packing and evolution. From this we can only conclude that if the above mentioned GPCRs do associate in oligomers, their association is likely to be weak.

Recently a structure of a class F GPCR, human Smoothened receptor [PDB: 4jkv], was solved [41] showing yet again the very well conserved 7-TM bundle. A possible dimer interface is also observed in the asymmetric unit involving helices IV and V. The structure was engineered fusing a BRIL protein N-terminally to the receptor, but BRIL does not participate in the interface. We analyzed the interface as before with the EPPIC software and find this time a very different picture than for any of the class A receptors above. In this instance the area buried in the interface is fairly large (1200 $Å^2$) and more importantly each side of the interface buries 4 residues thus counting a total of 8 core residues, a good indication of a biological interface. Moreover the evolutionary indicators both agree on assigning a biological character to the interface (see Table 1). Thus in contrast to those above, we would propose a valid dimerization interface for the human Smoothened receptor. In this case, supporting evidence from FRET experiments shows that the *Drosophila melanogaster* Smoothened receptor dimerizes [42] *in vivo*. The human and fly receptors share 43% sequence identity.

As an additional control for the class A GPCR analysis we analyzed the structure of the β2 adrenergic receptor complexed with G-protein [43], where a *bona fide* biological interface exists between the receptor and the G-protein. The interface has a larger area than most of those above (1200 $Å^2$) and more importantly buries 8 residues in total, typical of biological interfaces [6]. The evolutionary analysis by EPPIC shows also a very strong signal in both the core-rim and the core-surface indicators (see last entry of Table 1). It must be noted, however, that this interface, albeit a validated GPCR-partner protein interface, is not TM-spanning, which limits its value as a positive control.

Conclusions

We have carried out a comprehensive study of all known validated TM protein-protein interfaces with high resolution and good crystallographic quality. A dataset of biological protein-protein interfaces should serve the community by facilitating further studies on membrane protein oligomerization. While we are aware that the dataset represent a small sample of the membrane protein structure space and is not bias-free, we are convinced that it contains enough data to enable useful findings.

The TM protein interfaces we studied are in broad terms not very different from those of soluble proteins: intimate packing with buried residues is needed for stable TM interfaces to form. Furthermore the residues involved in the core of the oligomerization surfaces are mostly similar in character to those in soluble proteins interfaces with a clear preference for hydrophobic ones, though alanine and glycine are to some extent overrepresented in the TM interfaces.

Importantly we conclude from our evolutionary analysis that the fingerprint of evolution can be detected in TM interfaces almost as well as in their soluble counterparts. TM interfaces possess a core of well-conserved residues that can serve to identify them when comparing against the average selection pressure of the rim of the interfaces or of the rest of the protein surface.

Additionally, we could not find significant crystallographic evidence for lipids mediating protein-protein interfaces in the transmembrane region. It must also be noted that crystallography does not seem to be ideally suited for studying membrane lipids, as their electron density almost invariably appears incomplete due to high mobility and conformational flexibility.

We also studied the proposed class A GPCR dimerization interfaces in the literature through our EPPIC method, finding that none of them seems to be a stable biological interface in light of the geometrical and evolutionary analysis. We cannot however rule out that one or more of the analyzed interfaces is a weak/transient biological interface. The recent class F GPCR structure of the human Smoothened receptor does in contrast show a clear signature of a biological interface.

Methods

Compilation and annotation of new reference dataset

The MPSTRUC database from Stephen White's lab was downloaded in XML format on the 5th of October 2012. From the entries we kept those that were solved by X-ray crystallography of 3-dimensional crystals, resolution was better than 2.8 Å and Rfree below 30%. Within those constraints, we selected for further screening the best resolution representative of each cluster of identical proteins. That resulted in 69 structures from the beta class and 105 from the alpha class. We then did manual curation of each of the entries by checking the relevant literature, in order to find out whether their oligomerization state was well established and backed up by experimental data independent from crystallography. From those we could validate 3 beta monomers, 16 alpha monomers, 16 beta oligomers and 46 alpha oligomers. The 62 oligomers were then manually inspected in order to find out which of the interfaces were spanning the TM region. We checked the membrane location with the help of the OPM [44] and PDBTM [45] databases. Some of the interfaces spanned both the TM as well as the soluble regions. In those cases, interfaces that were mostly in the soluble regions were discarded.

Table S1 contains the full list of interfaces together with their buried areas and the EPPIC results for each of them. Table S3 contains the annotations and literature references with evidence of their oligomerization states.

Interface geometry and EPPIC analysis

Interfaces were calculated with the EPPIC package [6], using the default parameters: cofactors were considered as part of the protein surfaces for the ASA calculations whenever they were larger than 40 non-Hydrogen atoms. Interface core residues are considered those that bury more than 95% of their ASAs upon interface formation [26]. For the evolutionary predictions the version 2013_02 of the UniProt database was used. An evolutionary call could be given if at least 10 sequence homologs could be found within 60% identity of the query, or if not enough the identity cut-off was relaxed to 50%. In the evolutionary scores (core-rim and core-surface), the core residues are defined as those burying more than 70% of their ASAs upon interface formation as per EPPIC defaults.

Residue propensities and enrichment

Statistics were gathered for both our newly compiled biological TM interfaces dataset and several datasets of biological soluble interfaces: DCbio [6], PLP [26], Ponstingl dimers [46] and Bahadur dimers [47]. The enrichments are defined as the log-odds ratios of frequencies in interface core residues (at 95% burial cut-off) with respect to the frequencies of all residues in the full proteins. The size of the dots in Figure 2 corresponds to the averaged frequency of each of the amino acids in both soluble protein set and membrane protein set. All plots were done with the open-source R statistical package [48].

The amino acids were grouped as follows:

- Hydrophobic: Ile, Leu, Met, Phe, Trp, Tyr, Val
- Polar: Asn, Gln, Ser, Thr
- Charged: Arg, Asp, Glu, Lys
- Aliphatic: Ile, Leu, Val
- Aromatic: His, Phe, Trp, Tyr
- Small: Ala, Asn, Asp, Cys, Gly, Pro, Ser, Thr, Val
- Tiny: Ala, Gly, Ser

Lipid analysis

In order to find out lipids at interfaces the command line version of EPPIC was used and run with two different settings: 1) calculating BSAs ignoring all small molecules, 2) calculating BSAs taking molecules of more than 20 non-Hydrogen atoms as attached to their corresponding chains. Any change of interface area or interface core residues between the two runs was then inspected manually for possible lipid interactions at the interfaces.

For the Twilight analysis the version 2013-01-16 of the Twilight annotations was downloaded from the program server [31]. 11 representative PDB membrane protein structures were selected from the alpha subclass covering some of the most important groups of membrane proteins. Only those that contained some lipids and that were present in Twilight, which depends on the PDB entries being present in the EDS server [49], could be taken.

List of abbreviations

EPPIC: Evolutionary Protein Protein Interface Classifier; PDB: Protein Data Bank; GPCR: G-protein coupled receptor; ASA: Accessible Surface Area; BSA: Buried Surface Area. TMP: Trans Membrane Protein. TM : Trans Membrane

Acknowledgements

Financial support from the Research Committee of the Paul Scherrer Institut and the SNF to GC (grant numbers FK-05.08.1, FK-04.09.1 and 140879, respectively) is gratefully acknowledged.

Author contributions

References

1. Deisenhofer J, Epp O, Miki K, Huber R, Michel H: **Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3Å resolution**. *Nature* 1985, **318**:618–624.

2. Ulmschneider MB, Sansom MS: **Amino acid distributions in integral membrane protein structures.** *Biochimica et biophysica acta* 2001, **1512**:1–14.

3. Adamian L, Naveed H, Liang J: **Lipid-binding surfaces of membrane proteins: evidence from evolutionary and structural analysis.** *Biochimica et biophysica acta* 2011, **1808**:1092–102.

4. Adamian L, Liang J: **Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins.** *Journal of molecular biology* 2001, **311**:891–907.

5. Jackups R, Liang J: **Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction.** *Journal of molecular biology* 2005, **354**:979–93.

6. Duarte JM, Srebniak A, Schärer MA, Capitani G: **Protein interface classification by evolutionary analysis.** *BMC bioinformatics* 2012, **13**:334.

7. Palsdottir H, Hunte C: **Lipids in membrane protein structures.** *Biochimica et biophysica acta* 2004, **1666**:2–18.

8. Barrera NP, Zhou M, Robinson C V: **The role of lipids in defining membrane protein interactions: insights from mass spectrometry.** *Trends in cell biology* 2013, **23**:1–8.

9. Ernst AM, Contreras FX, Brügger B, Wieland F: **Determinants of specificity at the protein-lipid interface in membranes.** *FEBS letters* 2010, **584**:1713–20.

10. Periasamy A: **Fluorescence resonance energy transfer microscopy: a mini review.** *Journal of biomedical optics* 2001, **6**:287–91.

11. Gautier I, Tramier M, Durieux C, Coppey J, Pansu RB, Nicolas JC, Kemnitz K, Coppey-Moisan M: **Homo-FRET microscopy in living cells to measure monomer-dimer transition of GFP-tagged proteins.** *Biophysical journal* 2001, **80**:3000–8.

12. Haupts U, Tittor J, Oesterhelt D: **Closing in on bacteriorhodopsin: progress in understanding the molecule.** *Annual review of biophysics and biomolecular structure* 1999, **28**:367–99.

13. Brouillette CG, McMichens RB, Stern LJ, Khorana HG: **Structure and thermal stability of monomeric bacteriorhodopsin in mixed phospholipid/detergent micelles.** *Proteins* 1989, **5**:38–46.

14. Grzesiek S, Dencher NA: **Monomeric and aggregated bacteriorhodopsin: Single-turnover proton transport stoichiometry and photochemistry.** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85**:9509–13.

15. Faham S, Bowie JU: **Bicelle crystallization: a new method for crystallizing membrane proteins yields a monomeric bacteriorhodopsin structure.** *Journal of molecular biology* 2002, **316**:1–6.

16. Gurevich V V, Gurevich E V: **How and why do GPCRs dimerize?** *Trends in pharmacological sciences* 2008, **29**:234–40.

17. Chabre M, Le Maire M: **Monomeric G-protein-coupled receptor as a functional unit.** *Biochemistry* 2005, **44**:9395–403.

18. Park PS-H, Filipek S, Wells JW, Palczewski K: **Oligomerization of G protein-coupled receptors: past, present, and future.** *Biochemistry* 2004, **43**:15643–56.

19. Wu H, Wacker D, Mileni M, Katritch V, Han GW, Vardy E, Liu W, Thompson AA, Huang X-P, Carroll FI, Mascarella SW, Westkaemper RB, Mosier PD, Roth BL, Cherezov V, Stevens RC: **Structure of the human κ-opioid receptor in complex with JDTic.** *Nature* 2012, **485**:327–32.

20. Salom D, Lodowski DT, Stenkamp RE, Le Trong I, Golczak M, Jastrzebska B, Harris T, Ballesteros JA, Palczewski K: **Crystal structure of a photoactivated deprotonated intermediate of rhodopsin.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:16123–8.

21. Wu B, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V, Stevens RC: **Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists.** *Science (New York, N.Y.)* 2010, **330**:1066–71.

22. Huang J, Chen S, Zhang JJ, Huang X-Y: **Crystal structure of oligomeric β(1)-adrenergic G protein-coupled receptors in ligand-free basal state.** *Nature structural & molecular biology* 2013, **20**:419–25.

23. Bamber L, Harding M, Monné M, Slotboom D-J, Kunji ERS: **The yeast mitochondrial ADP/ATP carrier functions as a monomer in mitochondrial membranes.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:10830–4.

24. Bamber L, Harding M, Butler PJG, Kunji ERS: **Yeast mitochondrial ADP/ATP carriers are monomeric in detergents.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:16224–9.

25. Nury H, Dahout-Gonzalez C, Trézéguet V, Lauquin G, Brandolin G, Pebay-Peyroula E: **Structural basis for lipid-mediated interactions between mitochondrial ADP/ATP carrier monomers.** *FEBS letters* 2005, **579**:6031–6.

26. Schärer MA, Grütter MG, Capitani G: **CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts.** *Proteins* 2010, **78**:2707–2713.

27. Eilers M, Patel AB, Liu W, Smith SO: **Comparison of helix interactions in membrane and soluble alpha-bundle proteins.** *Biophysical journal* 2002, **82**:2720–36.

28. Hackenberg H, Klingenberg M: **Molecular weight and hydrodynamic parameters of the adenosine 5'-diphosphate--adenosine 5'-triphosphate carrier in Triton X-100.** *Biochemistry* 1980, **19**:548–55.

29. Pebay-Peyroula E, Dahout-Gonzalez C, Kahn R, Trézéguet V, Lauquin GJ-M, Brandolin G: **Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside.** *Nature* 2003, **426**:39–44.

30. Schobert B, Cupp-Vickery J, Hornak V, Smith S, Lanyi J: **Crystallographic structure of the K intermediate of bacteriorhodopsin: conservation of free energy after photoisomerization of the retinal.** *Journal of molecular biology* 2002, **321**:715–26.

31. Pozharski E, Weichenberger CX, Rupp B: **Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures.** *Acta crystallographica. Section D, Biological crystallography* 2013, **69**:150–67.

32. Li D, Lee J, Caffrey M: **Crystallizing Membrane Proteins in Lipidic Mesophases. A Host Lipid Screen.** *Crystal growth & design* 2011, **11**:530–537.

33. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Kuhn P, Weis WI, Kobilka BK, Stevens RC: **High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.** *Science (New York, N.Y.)* 2007, **318**:1258–65.

34. Gurevich V V, Gurevich E V: **GPCR monomers and oligomers: it takes all kinds.** *Trends in neurosciences* 2008, **31**:74–81.

35. Babcock GJ, Farzan M, Sodroski J: **Ligand-independent dimerization of CXCR4, a principal HIV-1 coreceptor.** *The Journal of biological chemistry* 2003, **278**:3378–85.

36. Percherancier Y, Berchiche YA, Slight I, Volkmer-Engert R, Tamamura H, Fujii N, Bouvier M, Heveker N: **Bioluminescence resonance energy transfer reveals ligand-induced conformational changes in CXCR4 homo- and heterodimers.** *The Journal of biological chemistry* 2005, **280**:9895–903.

37. Janin J, Rodier F: **Protein-protein interaction at crystal contacts.** *Proteins* 1995, **23**:580–7.

38. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server.** *Trends in biochemical sciences* 1998, **23**:358–61.

39. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *Journal of molecular biology* 2001, **313**:399–416.

40. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state.** *Journal of Molecular Biology* 2007, **372**:774–797.

41. Wang C, Wu H, Katritch V, Han GW, Huang X-P, Liu W, Siu FY, Roth BL, Cherezov V, Stevens RC: **Structure of the human smoothened receptor bound to an antitumour agent**. *Nature* 2013:1–8.

42. Zhao Y, Tong C, Jiang J: **Hedgehog regulates smoothened activity by inducing a conformational switch.** *Nature* 2007, **450**:252–8.

43. Rasmussen SGF, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah STA, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK: **Crystal structure of the β2 adrenergic receptor-Gs protein complex.** *Nature* 2011, **477**:549–55.

44. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI: **OPM: orientations of proteins in membranes database.** *Bioinformatics (Oxford, England)* 2006, **22**:623–5.

45. Tusnády GE, Dosztányi Z, Simon I: **Transmembrane proteins in the Protein Data Bank: identification and classification.** *Bioinformatics (Oxford, England)* 2004, **20**:2964–72.

46. Ponstingl H, Kabir T, Thornton JM: **Automatic inference of protein quaternary structure from crystals**. *Journal of Applied Crystallography* 2003, **36**:1116–1122.

47. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **Dissecting subunit interfaces in homodimeric proteins.** *Proteins* 2003, **53**:708–19.

48. Team RDC: **R: A Language and Environment for Statistical Computing**. *R Foundation for Statistical Computing Vienna Austria* 2010.

49. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA: **The Uppsala Electron-Density Server.** *Acta crystallographica. Section D, Biological crystallography* 2004, **60**:2240–9.

Figure legends

Figure 1

Frequencies of the different amino acids in both trans-membrane protein interfaces and soluble ones. The TM interfaces are further subdivided into alpha and beta classes. The inset in the top left is a magnification of the lower part of the plot.

Figure 2

Enrichments of amino acids in both trans-membrane protein interfaces and soluble protein interfaces. The size of the dots represents the averaged frequencies of amino acids in both soluble and membrane protein sets.

Figure 3

Frequencies for the different groups of amino acids in interface core residues for either interface set: soluble proteins and transmembrane proteins. See Methods for the amino acid properties grouping.

Table 1

PDB	Chains	BSA	Cores 95%	Cores 70%	# seqs.	core-rim	core-surface
2i35	A+A	316.6	0+0	4	105	0.36*	-0.29*
2i36	C+C	684.5	1+1		105	0.46	-0.37
	A+B	509.9	1+1	2	105	0.97*	0.40*
2i37	A+B	418.2	0+0	4	105	0.41*	-0.31*
	C+C	413.2	0+0	4	105	0.38*	-0.33*
3odu	A+B	1209.3(801.8)	0+0		44	1.34	1.81
3oe0	A+A	1089.4	0+0		71	1.64	1.84
3oe6	A+A	1037.6(764.8)	0+0		83	1.69	2.98
3oe8	B+C	665.2(591.9)	0+0	7	71	1.38*	1.74*
3oe9	A+B	959.4(877.4)	0+0		90	1.47	1.84
4djh	A+B	1024.0	1+1		102	1.34	1.06
4gpo	A+B	833.5	0+0		49	1.76	2.99

4jkv	A+B	1237.7	6+6	18	0.51	-1.25
3sn6	A+R	1263.1	2+6	116,124	0.35	-2.32

The analyzed GPCR interfaces: a set of class A GPCR dimer interfaces proposed in the literature plus the proposed dimer interface for the human Smoothened receptor [PDB: 4jkv] and the β2 adrenergic receptor to G-protein interface [PDB: 3sn6]. In cases where the T4L fusion protein contributes to the interface the areas with and without (in brackets) the fusion proteins are shown. The evolutionary scores of interfaces where not enough core residues at 70% burial were present are marked with a star.

Table S1

The full list of all validated TM protein-protein interfaces and the EPPIC values calculated for them. Id is the interface identifier, starting from 1 for the largest interface in crystal and higher ids for increasingly smaller interfaces. n1 and n2 are the number of homologs used to calculate evolutionary scores for each interface partner. If both are below 10, no evolutionary prediction can be made and thus a "nopred" appears for the evolutionary calls. In the evolutionary score fields (core-rim and core-surface) a few different issues (not shown) can lead to "nopred" calls, e.g. not enough core residues, too many mutations in core or rim with respect to wild type etc. Also NaNs will be present when no score can be calculated for a number of reasons. The final field contains the number of votes (each of the 3 indicators casts 1 vote) that lead to the final call. The value 0 votes means that the final call was based on applying a hard-area cut-off.

Table S2

Twilight values for all lipids of 11 representative TM proteins. The Real Space Correlation Coefficient is given (RSCC) and also the final Twilight assessment: Y if the molecule is below the RSCC=0.6 threshold and thus was a Twilight positive, i.e. very likely to be wrongly modelled; N if its RSCC is above the 0.6 threshold; G if the RSCC of the molecule is above the 0.95 threshold, indicating highly confident modelling.

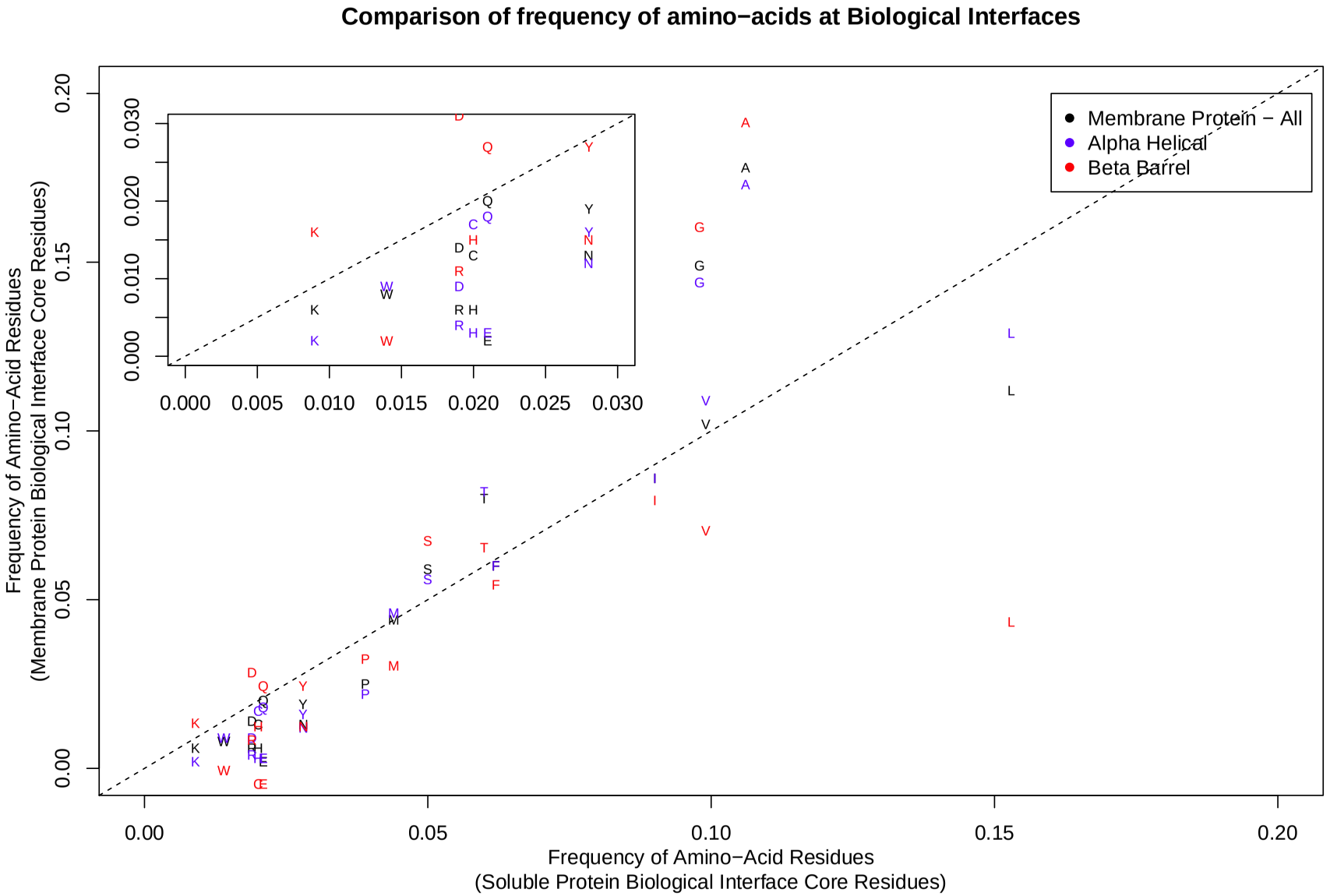
Table S3

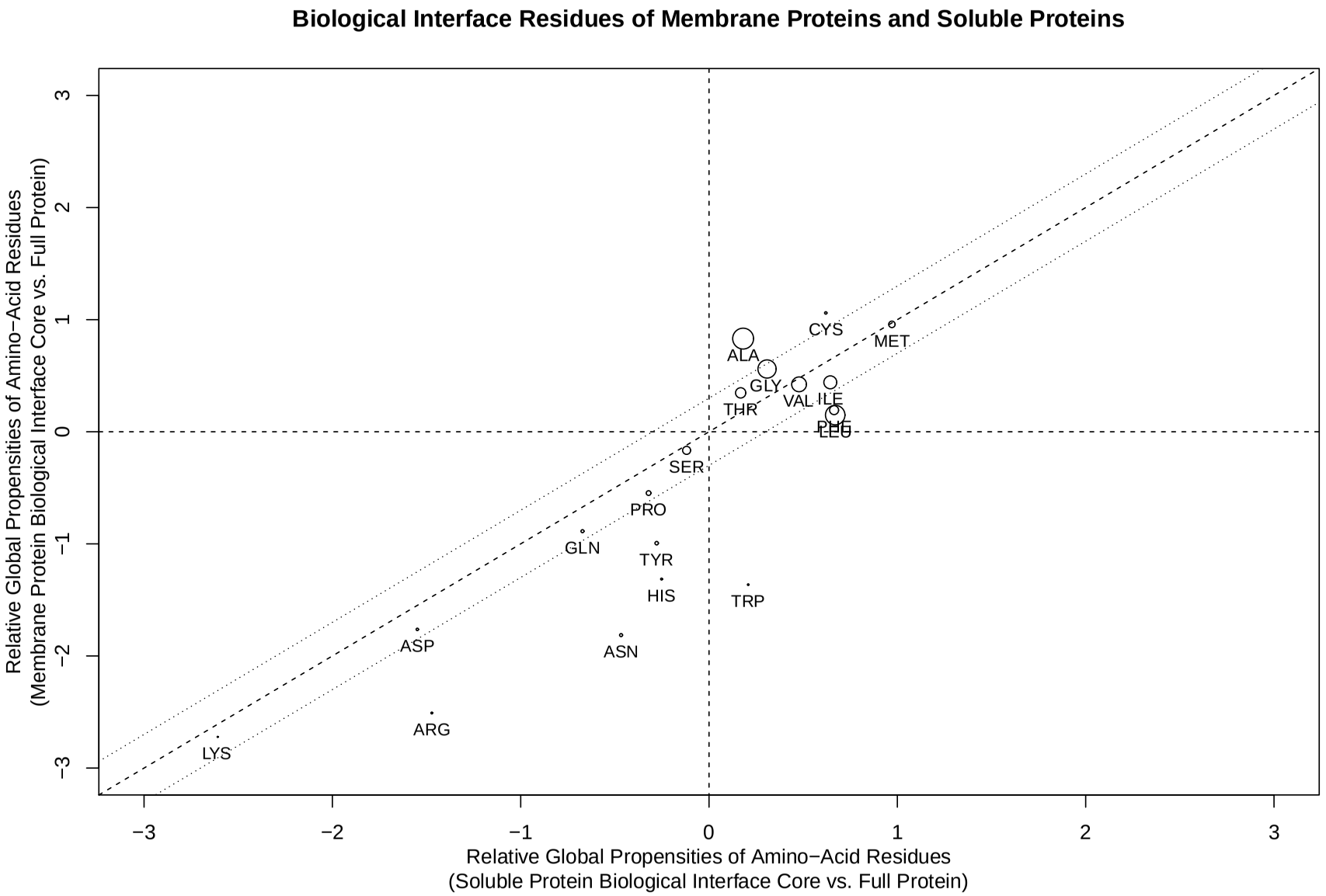
Manually curated TMPBio dataset, with experimental evidence from the literature. Columns are: PDB code, size of validated assembly (number of subunits), the list of all biological interfaces in the protein, the list of all TM biological interfaces in the protein (with a "*" if the interface spans both transmembrane and soluble regions), the Point Group symmetry, experimental technique used to verify the oligomeric state, reference where the evidence was found (given mostly as Pubmed links) and comment containing our annotation.

The table is additionally divided into subsections (with titles in bold in first columns) corresponding to the subdivisions present in Stephen's White MPSTRUC database. All the subsection titles have been kept even when no representative PDB structure for the section was found, either because of resolution criterium or because no structure could be validated as oligomer.

Experimental evidence abbreviations used: SEC size exclusion chromatography; AUC analytical gel filtration; AUC (SV) analytical ultracentrifugation sedimentation velocity; SLS, DLS, LS (static/dynamic) light scattering; MALS multi-angle light scattering; MALLS multi-angle laser light scattering; CCL chemical cross-linking; FRET fluorescence resonance energy transfer; NMR

nuclear magnetic resonance; SAXS small angle X-ray scattering; MS mass spectrometry; native-PAGE native polyacrylamide gel electrophoresis; AFM atomic force microscopy.


Comparison of frequency of amino−acids at Biological Interfaces
Membrane Protein − All
Alpha Helical
Beta Barrel
Frequency of Amino−Acid Residues
(Membrane Protein Biological Interface Core Residues)
Frequency of Amino−Acid Residues
(Soluble Protein Biological Interface Core Residues)
0.00
0.05
0.10
0.15
0.20
0.000
0.005
0.010
0.015
0.020
0.025
0.030


Biological Interface Residues of Membrane Proteins and Soluble Proteins
Relative Global Propensities of Amino−Acid Residues
(Membrane Protein Biological Interface Core vs. Full Protein)
Relative Global Propensities of Amino−Acid Residues
(Soluble Protein Biological Interface Core vs. Full Protein)
−3
−2
−1
0
1
2
3
CYS
MET
ALA
GLY
VAL
ILE
THR
PHE
LEU
SER
PRO
GLN
TYR
HIS
TRP
ASP
ASN
ARG
LYS

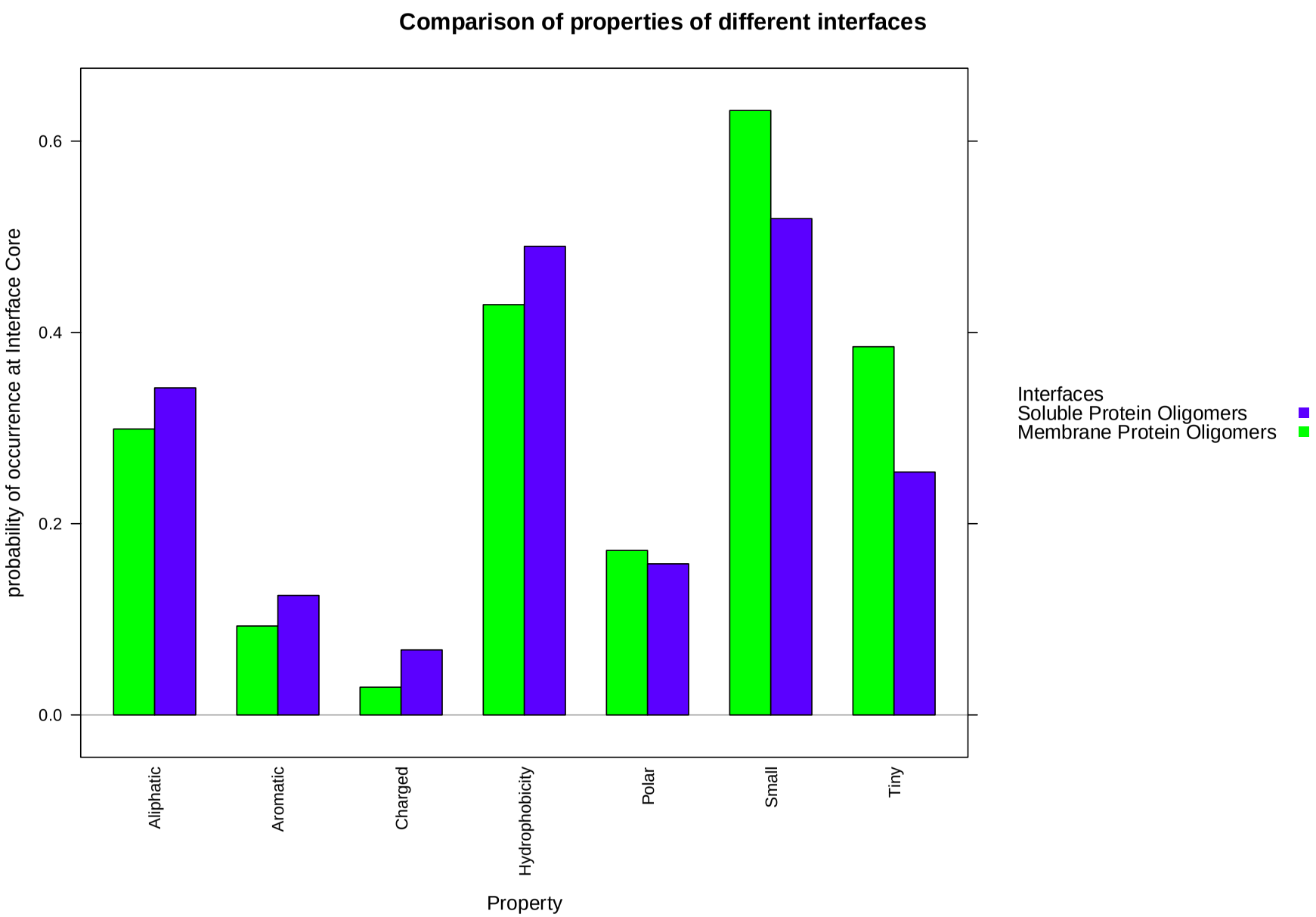

Comparison of properties of different interfaces
probability of occurrence at Interface Core
0.0
0.2
0.4
0.6
Aliphatic
Aromatic
Charged
Hydrophobicity
Polar
Small
Tiny
Property
Interfaces
Soluble Protein Oligomers
Membrane Protein Oligomers

Table S1

PDB	Id	BSA	n1	n2	geometry		core-rim		core-surface		final	
1a0t	1	1,967.37	55	55	23	bio	0.60	bio	-3.11	bio	3	bio
1a0t	2	1,965.86	55	55	23	bio	0.62	bio	-3.10	bio	3	bio
1a0t	3	1,964.43	55	55	23	bio	0.61	bio	-3.02	bio	3	bio
1af6	1	1,844.06	77	77	23	bio	0.60	bio	-2.76	bio	3	bio
1af6	2	1,842.20	77	77	22	bio	0.60	bio	-2.76	bio	3	bio
1af6	3	1,840.30	77	77	23	bio	0.60	bio	-2.75	bio	3	bio
1ek9	1	2,441.63	96	96	4	xtal	0.74	bio	-1.35	bio	0	bio
1ek9	2	2,438.95	96	96	4	xtal	0.69	bio	-1.44	bio	0	bio
1ek9	3	2,430.41	96	96	5	xtal	0.70	bio	-1.47	bio	0	bio
1eys	1	4,828.53	83	102	57	bio	0.68	bio	-7.74	bio	0	bio
1jb0	1	3,967.59	105	107	78	bio	0.38	bio	-5.86	bio	0	bio
1jb0	5	1,105.55	49	105	8	bio	0.66	bio	NaN	bio	3	bio
1jb0	7	915.10	22	22	4	xtal	0.75	bio	-1.51	bio	2	bio
1k4c	2	1,081.94	9	9	10	bio	0.73	nopred	-2.59	nopred	1	bio
1ldf	1	1,581.38	102	102	23	bio	0.46	bio	-2.19	bio	3	bio
1lgh	1	940.40	12	8	3	xtal	0.27	bio	-3.85	bio	2	bio
1lgh	3	939.70	12	8	3	xtal	0.27	bio	-3.91	bio	2	bio
1lgh	5	600.77	8	8	2	xtal	1.09	nopred	-0.60	nopred	1	xtal
1lgh	8	581.01	8	8	1	xtal	1.50	nopred	-0.07	nopred	1	xtal
1ppj	14	1,115.12	67	79	3	xtal	0.73	bio	-0.39	xtal	2	xtal
1q16	7	824.59	74	74	0	xtal	1.24	xtal	-0.41	xtal	3	xtal
1qd6	1	1,441.46	92	92	3	xtal	0.39	bio	-2.18	bio	2	bio
1rwt	2	995.20	119	119	10	bio	0.63	bio	-1.90	bio	3	bio
1rwt	3	979.02	119	119	9	bio	0.57	bio	-1.68	bio	3	bio
1rwt	5	962.82	119	119	9	bio	0.64	bio	-1.82	bio	3	bio
1u7g	1	1,707.86	90	90	22	bio	0.94	xtal	-1.04	bio	2	bio
1uun	1	1,892.53	29	29	9	bio	0.62	bio	-1.51	bio	3	bio
1uun	2	1,891.08	29	29	10	bio	0.56	bio	-1.60	bio	3	bio
1v54	4	2,716.39	124	46	42	bio	0.35	bio	-4.46	bio	0	bio
1v54	6	1,904.80	82	124	17	bio	0.59	bio	NaN	xtal	2	bio
1v54	8	1,810.38	102	46	21	bio	0.70	bio	NaN	xtal	2	bio
1v54	15	1,372.85	118	124	13	bio	0.73	bio	-1.38	bio	3	bio
1v54	30	910.99	46	82	6	bio	1.31	xtal	1.46	xtal	2	xtal
1v54	31	718.29	33	102	1	xtal	0.90	nopred	-1.51	nopred	1	xtal
1yc9	1	2,581.47	11	11	11	bio	0.47	bio	-1.70	bio	0	bio
1z98	1	1,749.51	118	118	26	bio	0.37	bio	-3.32	bio	3	bio
1z98	2	1,744.76	118	118	23	bio	0.40	bio	-3.52	bio	3	bio
2b2f	1	1,905.34	31	31	35	bio	0.84	xtal	-1.35	bio	2	bio
2bhw	1	915.74	119	119	7	bio	0.53	bio	-1.73	bio	3	bio
2bhw	2	915.19	119	119	8	bio	0.53	bio	-1.71	bio	3	bio
2bhw	3	910.88	119	119	7	bio	0.53	bio	-1.72	bio	3	bio
2bs2	3	2,050.12	16	16	8	bio	0.81	xtal	-1.08	bio	2	bio
2f2b	1	1,764.89	24	24	26	bio	0.63	bio	-1.13	bio	3	bio
2fgr	1	1,232.88	10	10	15	bio	0.35	bio	-3.24	bio	3	bio
2gr8	2	1,089.95	8	8	11	bio	1.04	nopred	1.66	nopred	1	bio
2gr8	3	1,087.53	8	8	9	bio	0.86	nopred	1.78	nopred	1	bio
2gr8	4	1,082.06	8	8	10	bio	0.93	nopred	2.05	nopred	1	bio
2j1n	1	1,570.40	110	110	18	bio	0.72	bio	-0.69	xtal	2	bio
2j1n	2	1,566.71	110	110	18	bio	0.78	xtal	-0.66	xtal	2	xtal
2j1n	3	1,564.96	110	110	18	bio	0.69	bio	-0.89	xtal	2	bio
2j58	1	4,294.95	107	107	15	bio	0.56	bio	-5.60	bio	0	bio
2j58	2	4,285.26	107	107	20	bio	0.51	bio	-5.82	bio	0	bio
2j58	3	4,257.43	107	107	20	bio	0.53	bio	-6.00	bio	0	bio
2j58	4	4,256.53	107	107	17	bio	0.52	bio	-5.69	bio	0	bio
2j58	5	4,255.17	107	107	20	bio	0.50	bio	-5.96	bio	0	bio
2j58	6	4,248.70	107	107	19	bio	0.55	bio	-6.04	bio	0	bio
2j58	7	4,242.90	107	107	19	bio	0.51	bio	-5.94	bio	0	bio
2j58	8	4,200.38	107	107	20	bio	0.51	bio	-5.82	bio	0	bio
2j7a	23	671.21	9	9	4	xtal	0.69	nopred	-0.14	nopred	1	xtal
2j8c	1	4,808.45	75	106	79	bio	0.52	bio	-7.99	bio	0	bio
2j8s	1	3,320.38	106	106	14	bio	0.35	bio	-2.85	bio	0	bio
2j8s	2	3,145.04	106	106	10	bio	0.32	bio	-2.76	bio	0	bio
2j8s	3	2,818.52	106	106	10	bio	0.56	bio	-2.20	bio	0	bio
2mpr	1	1,877.41	76	76	23	bio	0.57	bio	-2.74	bio	3	bio
2mpr	2	1,873.89	76	76	23	bio	0.54	bio	-2.84	bio	3	bio
2mpr	3	1,868.01	76	76	22	bio	0.54	bio	-2.82	bio	3	bio
2o4v	1	2,018.80	27	27	17	bio	0.76	xtal	-2.78	bio	2	bio
2o4v	2	2,015.21	27	27	17	bio	0.77	xtal	-2.75	bio	2	bio

continued on next page

Table S1 – continued from previous page

PDB	Id	BSA	n1	n2	geometry		core-rim		core-surface		final	
2o4v	3	1,996.42	27	27	16	bio	0.75	xtal	-2.78	bio	2	bio
2o9d	1	1,829.57	101	101	30	bio	0.46	bio	-2.52	bio	3	bio
2o9d	2	1,717.59	101	101	29	bio	0.45	bio	-2.50	bio	3	bio
2qi9	1	1,882.77	94	94	6	bio	0.60	bio	-3.11	bio	3	bio
2w2e	1	2,553.53	14	14	28	bio	0.46	bio	-0.94	xtal	0	bio
2wgm	5	1,924.29	77	77	31	bio	0.81	xtal	NaN	nopred	1	xtal
2wgm	8	1,922.22	77	77	31	bio	0.80	xtal	NaN	nopred	1	xtal
2wgm	11	1,919.28	77	77	30	bio	0.86	xtal	NaN	nopred	1	xtal
2wgm	19	1,915.75	77	77	31	bio	0.84	xtal	NaN	nopred	1	xtal
2wgm	20	1,915.47	77	77	32	bio	0.79	xtal	NaN	nopred	1	xtal
2wgm	21	1,914.82	77	77	31	bio	0.83	xtal	NaN	nopred	1	xtal
2wgm	22	1,914.72	77	77	32	bio	0.85	xtal	NaN	nopred	1	xtal
2wgm	33	1,900.78	77	77	30	bio	0.80	xtal	NaN	nopred	1	xtal
2wgm	35	1,897.38	77	77	31	bio	0.80	xtal	NaN	nopred	1	xtal
2wgm	38	1,895.48	77	77	31	bio	0.80	xtal	NaN	nopred	1	xtal
2wgm	44	1,887.79	77	77	30	bio	0.79	xtal	NaN	nopred	1	xtal
2wie	1	1,909.58	103	103	35	bio	1.19	xtal	NaN	nopred	1	xtal
2wie	2	1,889.23	103	103	33	bio	1.20	xtal	NaN	nopred	1	xtal
2wie	3	1,877.42	103	103	35	bio	1.22	xtal	NaN	nopred	1	xtal
2wie	4	1,867.64	103	103	34	bio	1.21	xtal	NaN	nopred	1	xtal
2wie	5	1,856.78	103	103	33	bio	1.36	xtal	NaN	nopred	1	xtal
2wjn	1	4,693.27	18	102	57	bio	0.79	xtal	-4.95	bio	0	bio
2wlj	1	2,247.77	9	9	13	bio	0.86	nopred	-1.22	nopred	0	bio
2wlj	2	2,183.44	9	9	10	bio	0.61	nopred	-1.90	nopred	1	bio
2wsw	2	715.72	23	23	3	xtal	0.45	bio	-0.33	xtal	2	xtal
2zfg	1	1,440.50	104	104	19	bio	0.78	xtal	-1.74	bio	2	bio
3arc	1	6,054.21	101	82	111	bio	0.80	xtal	NaN	nopred	0	bio
3arc	4	3,644.40	113	82	41	bio	2.05	xtal	NaN	xtal	0	bio
3arc	5	3,250.90	101	102	29	bio	1.15	xtal	NaN	bio	0	bio
3b9w	1	1,983.85	15	15	28	bio	0.49	bio	-4.20	bio	3	bio
3c02	1	1,591.30	6	6	22	bio	0.33	nopred	-1.82	nopred	1	bio
3cx5	6	1,491.21	28	104	5	xtal	1.03	xtal	-0.49	xtal	3	xtal
3d5k	1	2,739.74	102	102	7	bio	0.59	bio	-2.27	bio	0	bio
3d5k	2	2,738.93	102	102	10	bio	0.58	bio	-2.46	bio	0	bio
3d5k	3	2,726.55	102	102	8	bio	0.65	bio	-2.37	bio	0	bio
3d9s	1	1,558.88	87	87	22	bio	0.55	bio	-1.84	bio	3	bio
3d9s	2	1,547.05	87	87	21	bio	0.62	bio	-1.68	bio	3	bio
3d9s	3	1,542.02	87	87	22	bio	0.58	bio	-1.54	bio	3	bio
3d9s	4	1,527.44	87	87	24	bio	0.53	bio	-1.75	bio	3	bio
3gd8	1	1,515.83	35	35	16	bio	0.35	bio	-2.46	bio	3	bio
3hb3	1	3,887.81	30	106	26	bio	0.53	bio	-4.90	bio	0	bio
3jqo	1	2,342.02	10	10	4	xtal	0.83	xtal	1.78	xtal	0	bio
3jqo	2	2,336.49	10	10	5	xtal	0.85	xtal	NaN	xtal	0	bio
3jqo	3	2,327.38	10	10	5	xtal	0.84	xtal	NaN	xtal	0	bio
3jqo	4	2,323.32	10	10	6	bio	0.77	xtal	NaN	xtal	0	bio
3jqo	5	2,306.63	10	10	5	xtal	0.78	xtal	NaN	xtal	0	bio
3jqo	6	2,306.03	10	10	5	xtal	0.87	xtal	NaN	xtal	0	bio
3jqo	7	2,299.34	10	10	5	xtal	0.88	xtal	NaN	xtal	0	bio
3jqo	8	2,299.00	10	10	5	xtal	0.73	bio	NaN	xtal	0	bio
3jqo	9	2,291.00	10	10	5	xtal	0.77	xtal	NaN	xtal	0	bio
3jqo	10	2,288.86	10	10	5	xtal	0.81	xtal	2.37	xtal	0	bio
3jqo	11	2,280.57	10	10	5	xtal	0.79	xtal	NaN	xtal	0	bio
3jqo	12	2,276.64	10	10	5	xtal	0.82	xtal	1.90	xtal	0	bio
3jqo	13	2,273.28	10	10	5	xtal	0.91	xtal	NaN	xtal	0	bio
3jqo	14	2,227.60	10	10	5	xtal	0.79	xtal	1.02	xtal	0	bio
3k3f	1	1,689.32	0	0	12	bio	1,000.00	nopred	NaN	nopred	1	bio
3kcu	1	1,615.51	111	111	16	bio	0.69	bio	-1.87	bio	3	bio
3kcu	2	1,581.06	111	111	16	bio	0.56	bio	-2.03	bio	3	bio
3kcu	3	1,491.53	111	111	17	bio	0.57	bio	-2.10	bio	3	bio
3kcu	4	1,487.20	111	111	17	bio	0.57	bio	-2.10	bio	3	bio
3kcu	5	1,426.85	111	111	16	bio	0.54	bio	-2.17	bio	3	bio
3kly	1	1,689.28	34	34	16	bio	0.64	bio	-1.27	bio	3	bio
3kly	2	1,646.64	34	34	17	bio	0.58	bio	-1.41	bio	3	bio
3kly	3	1,637.08	34	34	16	bio	0.57	bio	-1.45	bio	3	bio
3kly	4	1,631.44	34	34	16	bio	0.58	bio	-1.35	bio	3	bio
3kly	5	1,616.54	34	34	17	bio	0.56	bio	-1.51	bio	3	bio
3ldc	1	821.44	3	3	5	xtal	0.30	nopred	-1.62	nopred	1	xtal
3m71	1	1,490.55	11	11	10	bio	0.64	bio	-1.62	bio	3	bio
3pik	1	2,662.31	33	33	4	xtal	0.82	xtal	-0.82	xtal	0	bio
3rlf	1	3,993.60	103	89	30	bio	0.31	bio	-5.97	bio	0	bio

continued on next page

Table S1 – continued from previous page

PDB	Id	BSA	n1	n2	geometry		core-rim		core-surface		final	
3tdo	1	1,711.47	15	15	23	bio	0.72	bio	-2.71	bio	3	bio
3tdo	2	1,610.37	15	15	24	bio	0.63	bio	-2.94	bio	3	bio
3tdo	3	1,603.49	15	15	24	bio	0.57	bio	-3.06	bio	3	bio
3tdo	4	1,595.40	15	15	22	bio	0.62	bio	-2.92	bio	3	bio
3tdo	5	1,581.25	15	15	22	bio	0.62	bio	-2.89	bio	3	bio
3tij	1	1,259.18	105	105	10	bio	0.51	bio	-2.27	bio	3	bio
3vzt	1	1,452.94	78	78	14	bio	0.43	bio	-2.51	bio	3	bio
4a01	1	3,265.69	111	111	61	bio	0.32	bio	-5.68	bio	0	bio
4av3	1	2,838.90	10	10	47	bio	0.56	bio	-2.83	bio	0	bio
4f4s	1	1,633.16	107	107	28	bio	0.98	xtal	NaN	nopred	1	xtal
4f4s	5	1,577.73	107	107	28	bio	0.87	xtal	NaN	nopred	1	xtal
4f4s	6	1,567.80	107	107	29	bio	0.84	xtal	NaN	nopred	1	xtal
4f4s	7	1,553.88	107	107	26	bio	1.00	xtal	NaN	nopred	1	xtal
4f4s	9	1,545.24	107	107	28	bio	0.82	xtal	NaN	nopred	1	xtal
7ahl	1	2,817.31	3	3	11	bio	0.96	nopred	0.14	nopred	0	bio
7ahl	2	2,809.31	3	3	11	bio	1.07	nopred	0.83	nopred	0	bio
7ahl	3	2,793.55	3	3	13	bio	0.89	nopred	0.14	nopred	0	bio
7ahl	4	2,789.40	3	3	12	bio	0.85	nopred	0.15	nopred	0	bio
7ahl	5	2,785.36	3	3	12	bio	0.88	nopred	0.05	nopred	0	bio
7ahl	6	2,763.94	3	3	12	bio	0.90	nopred	0.16	nopred	0	bio
7ahl	7	2,745.04	3	3	12	bio	1.04	nopred	0.50	nopred	0	bio

Table S2

PDB	Lipid name	Chain & Residue	RSCC	Resolution	Valid
3a7k	L3P	A 331	0.254	2.00	Y
3a7k	L2P	A 295	0.324	2.00	Y
3a7k	L3P	A 298	0.333	2.00	Y
3a7k	L3P	A 296	0.384	2.00	Y
3a7k	L3P	B 331	0.388	2.00	Y
3a7k	L3P	A 297	0.425	2.00	Y
3a7k	L2P	D 293	0.428	2.00	Y
3a7k	L3P	D 330	0.429	2.00	Y
3a7k	L3P	A 301	0.435	2.00	Y
3a7k	L2P	D 294	0.447	2.00	Y
3a7k	22B	D 300	0.456	2.00	Y
3a7k	L1P	B 293	0.463	2.00	Y
3a7k	L3P	A 299	0.511	2.00	Y
3a7k	L2P	B 295	0.512	2.00	Y
3a7k	L1P	B 294	0.515	2.00	Y
3a7k	L1P	A 293	0.547	2.00	Y
3a7k	L3P	A 333	0.548	2.00	Y
3a7k	22B	A 300	0.595	2.00	Y
3a7k	L2P	B 296	0.644	2.00	N
3a7k	L2P	A 294	0.727	2.00	N
3a7k	22B	B 300	0.819	2.00	N
1xio	PEE	A 310	0.690	2.00	N
1xio	PEE	A 311	0.710	2.00	N
1xio	PEE	A 307	0.711	2.00	N
1xio	PEE	A 312	0.717	2.00	N
1xio	PEE	A 313	0.735	2.00	N
1xio	PEE	A 309	0.741	2.00	N
1xio	PEE	A 315	0.756	2.00	N
1xio	PEE	A 306	0.788	2.00	N
1xio	PEE	A 304	0.796	2.00	N
1xio	PEE	A 303	0.799	2.00	N
1xio	PEE	A 302	0.839	2.00	N
1xio	PEE	A 305	0.850	2.00	N
1xio	PEE	A 308	0.857	2.00	N
2ei4	L2P	A 280	0.606	2.10	N
2ei4	22B	A 270	0.618	2.10	N
3ddl	PX4	B1415	0.638	1.90	N
3ddl	PCW	B1416	0.642	1.90	N
3ddl	SXN	A1401	0.689	1.90	N
3ddl	SXN	B1401	0.792	1.90	N
2z73	PC1	B1004	0.836	2.50	N
1k4c	DGA	C1001	0.639	2.00	N
1okc	PC1	A 983	0.517	2.20	Y
1okc	CDL	A 802	0.614	2.20	N
1okc	PC1	A 980	0.792	2.20	N
1okc	PC1	A 981	0.807	2.20	N
1okc	CDL	A 801	0.843	2.20	N
1okc	PC1	A 982	0.859	2.20	N
1okc	CDL	A 800	0.874	2.20	N
2c3e	CDL	A 802	0.654	2.80	N
2c3e	CDL	A 800	0.819	2.80	N
2c3e	CDL	A 801	0.898	2.80	N
1ppj	PEE	D2006	0.676	2.10	N
1ppj	CDL	P3003	0.798	2.10	N
1ppj	CDL	D2003	0.812	2.10	N
1ppj	PEE	Q3006	0.857	2.10	N
1ppj	CDL	G2004	0.861	2.10	N
1ppj	CDL	T3004	0.889	2.10	N
1ppj	PEE	P3007	0.916	2.10	N
1ppj	PEE	C2007	0.940	2.10	N
3arc	DGD	d 755	0.360	1.90	Y
3arc	DGD	D 755	0.388	1.90	Y
3arc	LMG	Z 784	0.482	1.90	Y
3arc	LMG	z 784	0.540	1.90	Y
3arc	LMG	C 776	0.558	1.90	Y
3arc	LHG	e 772	0.610	1.90	N
3arc	SQD	B 668	0.613	1.90	N
3arc	SQD	d 768	0.623	1.90	N

continued on next page

Table S2 – continued from previous page

pdb	lipid name	chain & residue	RSCC	resolution	valid
3arc	LMG	c 776	0.646	1.90	N
3arc	SQD	a 667	0.682	1.90	N
3arc	LHG	E 772	0.701	1.90	N
3arc	SQD	L 668	0.707	1.90	N
3arc	SQD	A 667	0.736	1.90	N
3arc	SQD	D 768	0.755	1.90	N
3arc	LMG	a 751	0.772	1.90	N
3arc	LMG	A 751	0.786	1.90	N
3arc	LMG	d 692	0.860	1.90	N
3arc	LMG	c 729	0.868	1.90	N
3arc	LMG	C 729	0.870	1.90	N
3arc	LMG	D 692	0.893	1.90	N
3arc	LMG	B 669	0.892	1.90	N
3arc	SQD	a 659	0.907	1.90	N
3arc	SQD	A 659	0.905	1.90	N
3arc	DGD	c 661	0.910	1.90	N
3arc	DGD	c 660	0.910	1.90	N
3arc	DGD	C 660	0.920	1.90	N
3arc	LMG	b 669	0.926	1.90	N
3arc	LHG	D 714	0.924	1.90	N
3arc	LHG	d 714	0.924	1.90	N
3arc	DGD	H 663	0.927	1.90	N
3arc	DGD	h 663	0.931	1.90	N
3arc	LHG	l 694	0.942	1.90	N
3arc	DGD	C 657	0.944	1.90	N
3arc	LHG	D 664	0.943	1.90	N
3arc	DGD	C 661	0.947	1.90	N
3arc	DGD	c 657	0.946	1.90	N
3arc	LHG	d 664	0.952	1.90	G
3arc	LHG	L 694	0.955	1.90	G
3arc	LHG	d 702	0.960	1.90	G
3arc	LHG	D 702	0.972	1.90	G
1rwt	DGD	G9632	0.883	2.72	N
1rwt	DGD	A 632	0.882	2.72	N
1rwt	DGD	B2632	0.889	2.72	N
1rwt	DGD	D5632	0.889	2.72	N
1rwt	DGD	H6632	0.890	2.72	N
1rwt	DGD	H7632	0.890	2.72	N
1rwt	DGD	E4632	0.897	2.72	N
1rwt	DGD	I8632	0.897	2.72	N
1rwt	DGD	B1632	0.902	2.72	N
1rwt	DGD	D3632	0.902	2.72	N
1rwt	LHG	A 630	0.923	2.72	N
1rwt	LHG	G6630	0.925	2.72	N
1rwt	LHG	B1630	0.928	2.72	N
1rwt	LHG	D3630	0.930	2.72	N
1rwt	LHG	E4630	0.930	2.72	N
1rwt	LHG	J9630	0.931	2.72	N
1rwt	LHG	H7630	0.935	2.72	N
1rwt	LHG	C2630	0.933	2.72	N
1rwt	LHG	F5630	0.935	2.72	N
1rwt	LHG	I8630	0.934	2.72	N

Table S3 - Dataset of validated TM protein interfaces

TRANSMEMBRANE PROTEINS: ALPHA-HELICAL

pdb	name	size	bio interfaces	bio TM interfaces	PG	evidence	reference	comments
	Adventitious Membrane Proteins: Alpha-helical Pore-forming Toxins.							No structures from this class could be validated.
	Outer Membrane Proteins							
2J58	Wza translocon for capsular polysaccharides	8	1-8	1-8*	C8	SDS-PAGE	17086202	A well-known octamer, SDS-stable (mentioned in the paper without the reference, seems to be an established fact). Note that the transmembrane region is formed by 8 times the C-terminal domain. The interfaces in any case go through all 3 other domains (periplasmic)
3JQO	Type IV outer membrane secretion complex	42	1-14,15-28, 29-42, 43-70	1-14*	C14	EM	19946264	A massive structure with C14 symmetry, unlikely to happen by chance in crystal and not in physiological conditions. The 14-fold assembly fits very well to the known EM maps. There is little doubt that this is biological. Interestingly this is a heterotetradecamer: a heterotrimer (proteins TraF, TraO and TraN) is repeated 14 times in C14 symmetry to assemble the pore. 1-14 interfaces are the heterologous interface of the main protein TraF C14 assembly, 15-28 the interface in the heterotrimer between TraF and TraO, 29-42 the interface in heterotrimer between TraO and TraN. 43-70 are more interfaces in AU that most likely are induced. Beyond 70 there's even more interfaces but smaller than 400A2, so we discard (as we've done in other cases). Interfaces 1-14 contain the TM part and also a periplasmic part, the rest of interfaces are not TM
	Bacterial and Algal Rhodopsins							No structures from this class could be validated.
	G Protein-Coupled Receptors (GPCRs)							No structures from this class could be validated.
	Autonomously Folding "Membrane Proteins" (Sec-independent)							No structures from this class could be validated.
	Virus Coat Proteins							No structures from this class could be validated.
	Glycoproteins							No structures from this class could be validated.
	Epidermal Growth Factor Receptors							No structures from this class could be validated.
	Erythropoietin-Producing Hepatocellular Receptors							No structures from this class could be validated.
	Integrin Adhesion Receptors							No structures from this class could be validated.
	Histidine Kinase Receptors							No structures from this class could be validated.
	Immune Receptors							No structures from this class could be validated.
	SNARE Protein Family							No structures from this class could be validated.
	Channels: Potassium and Sodium Ion-Selective							
1K4C	KcsA Potassium channel, H+ gated	4	2	2	C4	Inhibition model, geometry	1706481, 11689936 11689935	Bacterial K+ channel (Streptomyces Lividans). It seems to be a tetramer. The 1991 paper 1706481 predicts a tetramer with a complicated experiment using a known toxin inhibitor of the channel (from Drosophila) and a mutant of the channel known not to be inhibited. By modelling the system, the stoichiometry of the channel is calculated to be 4. Much later the structure came (2001) and its C4 geometry (tight helix cone formed in a 4-fold) is unlikely to have formed by chance in crystal. This 1k4c structure is crystallised with Fab fragments to obtain high resolution, a lower resolution (2.8) one without Fab is 1jvm. Both are different crystals and still conserve perfectly the same tetrameric helix cone, providing further evidence. Prokaryot and Eukaryot K channels seem to be very closely related (see 9525854)
3LDC	MthK Potassium channel, Ca++ gated	4	1	1	C4	SDS-PAGE	12037559	Archaeal K+ channel (Methanothermobacter thermautotrophicus), only transmembrane domain at high resolution. As with all other K+ channels well known to be a tetramer. The full length is seen as tetramer in SDS-PAGE (12037559). Note that the full length structure (1lnq) is quite low resolution. Also it is not clear enough whether it can be interpreted as they do in the paper. For those 2 reasons we don't use the full length for this data set, but only the high res transmembrane domain. The cytoplasmic gating

								domain has been also solved separately (2aef).
2WLJ	KirBac3.1 Inward-Rectifier Potassium channel (semi-latched)	4	1,2	1,2	C4	AFM	17936299, 16216578	The authors in first ref did AFM studies and found KirBac3.1 to be a tetramer formed by a dimer-of-dimers. This result was also found by authors in second ref who did cryo-EM on KirBac3.1
	Channels: Other Ion Channels							
3M71	SLAC1 anion channel, TehA homolog (wild-type)	3	1	1	C3	SEC-MALS, CCL	20981093	Trimer by both SEC-MALS and chemical cross linking. It's a structural genomics structure: a bacterial homolog (H. infuenzae) of a plant SLAC1 anion channel.
	Cys-Loop Receptor Family							No structures from this class could be validated.
	Channels: Aquaporins and Glyceroporins						11143978 10698922	All aquaporins are structurally similar and have a tetrameric assembly. They all occur as tetramers and have a rmsd of 0.5-2.5 A.
3GD8	AQP4 aquaporin water channel	4	1	1	C4	Fluorescence experiments	20071343, 7615928	Freeze free EM studies in order to study OAP association (Orthogonal Arrays of Particles, i.e. supramolecular assemblies beyond the tetramer) were reviewed by authors in second ref. Authors in first ref made Green fluorescent protein (GFP)-labeled M1 and M23 isoforms of AQP4 to address questions about AQP4 associations and OAP dynamics than cannot be addressed by available freeze-fracture electron microscopy, biochemical (native gel electrophoresis), or biophysical (single-particle tracking) methods. They also find that the 2 isoforms (differing only in a N-terminal tail) associate in heterotetramers by doing fluorescence experiments with the GFP labelled forms. They don't provide direct evidence for the homotetramer but this evidence should be enough as they are very similar.
3D9S	AQP5 aquaporin water channel (HsAQP5)	4	1-4	1-4	C4	Homology, Geometry		40% sequence id to 3gd8 above and very close structurally, plus C4 symmetry: we can safely call this a tetramer
2F2B	AqpM aquaporin water channel	4	1	1	C4	SDS-PAGE	12519768	An archaeal aquaporin. The authors here did SDS-PAGE analysis on AqpM and found that AqpM remained functional after incubations at temperatures above 80 °C and formed SDS-stable tetramers.
2O9D	AqpZ aquaporin water channel	4	1,2	1,2	C4	EM, SDS-PAGE	10518952 10518953, 9468603	AqpZ tetramer is stable in 1% SDS and runs as tetramers. EM studies also reveal tetrameric structure of AqpZ.
1Z98	SoPIP2;1 plant aquaporin (closed conformation)	4	1,2	1,2	C4	Similarity to other AQPs, Geometry		An aquaporin from spinach. Usual C4 symmetry and typical very conserved structure of aquaporins.
1LDF	GlpF glycerol facilitator channel	4	1	1	C4	EM	11265760	Negative stain electron microscopy of solubilized GlpF protein revealed a tetrameric structure of approximately 80 A side length. Scanning transmission electron microscopy yielded a mass of 170 kDa, corresponding to the tetrameric nature of GlpF.
3C02	PfAQP aquaglyceroporin	4	1	1	C4	Homology, Geometry		Homology to the other aquaporins plus usual C4 symmetry
2W2E	Aqy1 yeast aquaporin (pH 3.5)	4	1	1	C4	Homology, Geometry		Homology to the other aquaporins plus usual C4 symmetry
	Channels : Formate/Nitrite Transporter (FNT) Family							
3KCU	FocA, pentameric aquaporin-like formate transporter	5	1-5	1-5	C5	SEC-MALS, EM	19940917 20010838	FocA from E coli was crystallized by the authors in first ref. They get a pentameric assembly. But they do not do nay experiments to prove the pentameric assembly. Characterization of FocA from *Vibrio cholerae* (53% seq id, pentamers superpose with rmsd below 0.6) was done by authors in second ref. They did analytical size-exclusion chromatography coupled with static light scattering and refractive index techniques to get the molecular weight which corresponds to pentamer. Also the EM images suggest a pentamer.
3KLY	FocA formate transporter without formate	5	1-5	1-5	C5	SEC-MALS, EM	20010838	Characterization of FocA from *Vibrio cholerae* was done by authors. They did analytical size-exclusion chromatography coupled with static light scattering and refractive index techniques to get the molecular weight which corresponds to pentamer. Also the EM images suggest a pentamer.
3TDO	FNT3 Hydrosulphide Channel (HSC), pH 9.0	5	1-5	1-5	C5	Homology, Geometry	22407320	The structure of this HSC is also very similar to the FocA (rmsd 1.1). HSC also has a pentameric C5 assembly.
	Channels: Urea Transporters							
3K3F	Urea transporter	3	1	1	C3	CCL	19865084	The trimeric state of Urea transporter from *Desulfovibrio vulgaris* (dvUT) was studied by the authors with chemical crosslinking experiments. Purified dvUT was incubated with the amine-to-amine crosslinking agent disuccinimidyl glutarate at concentrations varying from 0-10 mM and then run on an SDS-PAGE gel. The peaks corresponds to the trimer of dvUT. Also, the same homotrimer was observed in a lower resolution structure obtained from the native protein, which crystallizes in a lower symmetry space group with different packing.
	Channels: Amt/Rh proteins							
1U7G	AmtB ammonia channel (mutant)	3	1	1	C3	SDS-PAGE, DLS, AUC, SEC	12023896	The authors here examined the quaternary structure of AmtB by SDS-PAGE, gel-filtration chromatography, dynamic light scattering and sedimentation ultracentrifugation: "The protein was resistant to dissociation by SDS and behaved as a stable oligomer on SDS-PAGE. By equilibrium desorption chromatography we determined the mass ratio of dodecyl b-D-maltoside to

								AmtB in the detergent-solubilized complex to be 1.03±0.03, and this allowed us to calculate, from analytical- ultracentrifugation data, that AmtB purifies as a trimer"
2B2F	Amt-1 ammonium channel	3	1	1	C3	Homology	16214888	Amt-1 from *Archaeglobus fulgidus* was crystallized by the authors. They found trimers in ASU. The structure of Amt-1 is very similar (rmsd ~1A) to that of AmtB from E coli. which is well characterized to be a trimer. So, we take Amt-1 also as a trimer by homology.
3B9W	Rh protein, possible ammonia or CO2 channel	3	1	1	C3	Homology	18032606, 16281947	The authors present a crystal structure of NeRh50 at 1.3 Å resolution from *Nitrosomonas europaea*. Homology modelling by authors in second ref shows that Rh is highly homologous to AmbT from E coli. which is well characterized to be a trimer. The structures are also similar. Thus we conclude by homology that NeRh is a trimer.
	Channels: Gap Junctions							No structures from this class could be validated.
	Channels: Intercellular							No structures from this class could be validated.
	Sec Proteins							No structures from this class could be validated.
	Oligosaccharyl-transferases (OST)							No structures from this class could be validated.
	Intramembrane Proteases							No structures from this class could be validated.
	Membrane-Bound Metalloproteases							No structures from this class could be validated.
	H+/Cl-Exchange Transporters							No structures from this class could be validated.
	CorA Superfamily Ion Transporters							No structures from this class could be validated.
	Bacterial Mercury Detoxification Proteins							No structures from this class could be validated.
	Multi-Drug Efflux Transporters							
2J8S	AcrB bacterial multi-drug efflux transporter (E. coli)	3	1,2,3	1*, 2*, 3*	C3	AUC-SV	17194213	This 2.54A resolution AcrB was not present in Blanco db. Added by us to have a representative of the AcrB cluster (which is in the db with a few other lower resolution members). It's a clear C3 trimer, they did AUC-SV. 2 darpins are attached to the trimer Interfaces 1,2,3 have a TM part and a soluble part.
	Membrane-Associated Proteins in Eicosanoid and Glutathione Metabolism (MAPEG)							No structures from this class could be validated.
	Major Facilitator Superfamily (MFS) Transporters							No structures from this class could be validated.
	Solute Sodium Symporter (SSS) Family							No structures from this class could be validated.
	Concentrative Nucleoside Transporters (CNT)							
3TIJ	Concentrative nucleoside transporter in complex with uridine	3	1	1	C3	CCL	22407322	To test whether vcCNT is trimeric, the authors performed structure-guided disulfide bridge cross-linking experiments. The cysteine mutants readily formed disulfide cross-linked trimers in both detergent micelles and cell membranes under oxidizing conditions.
	Nucleobase- Cation-Symport-1 (NCS1) Family							No structures from this class could be validated.
	Nucleobase- Cation-Symport-2 (NCS2) Family							No structures from this class could be validated.
	Betaine/ Choline/ Carnitine Transporter (BCCT) Family							
2WSW	CaiT carnitine transporter	3	2	2	C3	BN-PAGE, SEC, CCL, EM	20829798, 16365043	The pdb entry here corresponds to CaiT from Proteus Mirabilis and was crystallized by the authors in the first ref. CaiT from Proteus Mirabilis is very similar structurally and by sequence (~70% seq id) to CaiT from E. coli. Authors in second ref studies the oligomeric state of CaiT from E coli. Blue native gel electrophoresis indicated that CaiT was a

								trimer in detergent solution. This result was further supported by gel filtration, cross-linking and EM studies.
	Amino Acid/ Polyamine/ Organocation (APC) Superfamily							No structures from this class could be validated.
	Amino Acid Secondary Transporters							No structures from this class could be validated.
	Cation Diffusion Facilitator (CDF) Family							
	Antiporters							
1OKC	Mitochondrial ADP/ATP Carrier	1	-	-		SEC, Negative dominance studies	6243949, 14603310, 14498831, 16226253, 17056710, 17566106	ADP/ATP carrier from beef heart mitochondria was characterized by authors in first ref. They found that triton-solubilized CAT-protein complex exists as a dimer composed of two peptide subunits. The authors in second ref crystallized the ADP/ATP carrier from bovine heart.Apparently the inhibitor atractyloside produces a difference in dimerization (see third ref): with the inhibitor the structure is homodimer, without it is monomer (both by SEC). 1okc here does have the atractyloside inhibitor, thus it should be a dimer. BUT all interfaces are either in upside-down or head to tail orientations. Thus unlikely that any of those are real. 2c3e is the same protein (with atractyloside inhibitor) in a different crystal form solved later. Both do have an interface in common: 3 of 1okc and 1 of 2c3e, but extremely small (~290A2) and in an upside-down orientation with respect to each other. Interface 3 of 2c3e does have a 2 fold parallel orientation, with 219A2 of area. They claim in paper (4th ref) that it could be the interface that is compatible with the known homodimerization, but they are not convinced themselves as they say more evidence is needed. Later (ref 5) it was proven to be a monomer in detergent (by comprehensive SEC study and other techniques) and finally (ref 6) that it functions as a monomer in the membrane (by negative dominance studies). The case is settled for monomer.
	Apical Sodium-Dependent Bile Acid Transporters (ASBT)							No structures from this class could be validated.
	Energy-Coupling Factor (ECF) Transporters							
3RLB	ThiT, S component of the Thiamin Transporter	1	-	-		SEC-MALLS	20218726, 21706007	The authors in first ref characterized ThiT using Size exclusion chromatography coupled to static light scattering, refractive index, and UV absorbance measurements (SEC-MALLS). The molecular mass of ThiT in the DM micelle determined by the light scattering analysis was 22.7 kDa. As the molecular mass of ThiT calculated from the amino acid sequence was 21.2 kDa, it was concluded that ThiT was monomeric in the DM solubilized state.
	ATP Binding Cassette (ABC) Transporters							
2QI9	BtuCD Vitamin B12 Transporter	5	1-6	1		EPR	17673622	The authors here solve the crystal structure of BtuCD in complex with BtuF. BtuF resides in the periplasmic region. But the BtuCD-F complex has substantial conformational changes as compared with the previously reported structures of BtuCD and BtuF. By the Electron paramagnetic resonance (EPR) spectra studies done by the authors we can conclude that BtuCD-F chain contacts are biological. There are a total of five chains in the complex. 1 to 6 bio interfaces but only 1 is bio TM
3RLF	MalFGK2-MBP Maltose uptake transporter complex	5	1-6	1		SEC, CCL	21825153, 2026607	The authors in the first ref crystallized periplasmic MBP in complex with MalF, MalG and dimer of MalK from E coli. This complex was characterized by authors in second ref. They found that in all experiments, the MalF, MalG, and MalK proteins behaved as a multiprotein complex. They performed gel filtration experiment and Chemical cross-linking experimetns. Each complex contains two MalK, one MalF, and one MalG proteins. 1 to 6 bio interfaces but only 1 is bio TM
	Methyltransferases							No structures from this class could be validated.
	Phosphoenolpyruvate-Dependent Phosphotransferases (PTSs)							No structures from this class could be validated.
	Superfamily of K+ Transporters (SKT proteins)							No structures from this class could be validated.
	Membrane-Integral Pyrophosphatases (M-PPases)							
4A01	H+-translocating M-PPase	2	1	1	C2	RI, SEC, SDS-PAGE	10748246	The authors in the ref have listed all the methods and the respective references of the studies. The H'- PPase has been strongly suggested to exist as a dimer, using Radiation Inactivation, gel permeation HPLC and SDS/PAGE analysis. Although it is not clear whether higer oligomerization is also possible. We'll take at least the dimer interface to be biological.

4AV3	Na+-translocating M-PPase with metal ions in active site	2	1	1	C2	SEC-MALLS	21664973	This is a bacterial (Thermotoga Maritima) homolog of plant (mung bean) M-PPase above (4a01) with 37% seq id. Structural conservation between the 2 is very good and also for the dimer interface providing first strong clue for dimer. The authors in first ref perform Size exclusion chromatography coupled with static light scattering to proof that a few M-PPase proteins from different organisms are dimeric, including one from Symbiobacterium thermophilum which is 42% seq id from this one. Seems to be safe to call it a dimer.
	Bacterial V-type ATPase							No structures from this class could be validated.
	F-type ATPase							
4F4S	ATP synthase (F1c10)	10	1,5,6,7,9	1, 5, 6, 7, 9	C10	CCL	10576729, 9792682, 9642286	This is the very well studied subunit c of TM domain FO of ATP synthase from yeast, a C10 ring is seen in the crystal structure (high symmetry unlikely to be an artifact). The E coli homolog was well studied with cross linking and found to have 9-12 subunits assembling in a ring in the membrane, see refs 2 and 3. Additionally another crystal form was solved for the yeast ring (3u2f, 3u2y, 3u32, 3ud0) containing the exact same C10 ring. Also the full complex of F1 together with this c10 of FO was solved (1qo1, ref 1) and shows the 10 stoichiometry. There seems to be enough evidence for this assembly to be biological. The AU has 2 different half-barrels of the C10 assembly. Chains A-E are one half and chains K-O are the other. The C10 happens then on a 2 fold that completes the barrel by repeating A-E or K-O. We chose as bio TM interfaces those in the AU between K-O chains (interfaces 1,5,7,9) and the 2-fold interface between O and K (interface 6)
2WGM	Rotor (c11) of Na+-dependent F-ATP Synthase	11	5, 8, 11, 19, 20, 21, 22, 33, 35, 38, 44	5, 8, 11, 19, 20, 21, 22, 33, 35, 38, 44	C11	CCL, Homology, SDS-PAGE	9642181, 15860619, 19500592	The authors in first reference do a SPS-PAGE analysis on F1F0 ATPase from Ilyobacter tartaricus and finds that the c subunits forms a strong aggregate with the apparent molecular mass of 50 kDa which requires treatment with trichloroacetic acid for dissociation. That does not correspond to the 11x10KDa assembly seen in crystal but in any case is an indication that the protein oligomerizes very tightly. The authors in second ref crystallize F0 subdomain of ATPase from Ilyobacter tartaricus and solve the crystal structure. There are 4x11-mers barrels in AU, with 2 groups of 2 stacked vertically (unlikely to be natural). Other organisms seem to have similar same ring arrangements with different numbers of subunits (see review 11893513). In any case it seems to be an established fact that this c subunit is a C11 barrel. See for instance the CCL evidence for the yeast ATP synthase above.
2WIE	Rotor (c15) of H+-dependent F-ATP Synthase of an alkaliphilic cyanobacterium	15	1,2,3,4,5	1, 2, 3, 4, 5	C15	SDS-PAGE, AFM	16170308	The oligomeric c ring of the F-ATP synthase from the alkaliphilic cyanobacterium *Spirulina platensis* was studied by the authors in this ref. They found by SDS-PAGE experiments that the enzyme moved slowly than c14 assembly suggesting that the ring contains more than 14 subunits. AFM pictures from the same authors shows that the ring contains 15 subunits.
	P-type ATPase							No structures from this class could be validated.
	Phosphotransferases							No structures from this class could be validated.
	Hydrolases							No structures from this class could be validated.
	Oxygenases							No structures from this class could be validated.
	Oxidoreductases							
1Q16	NarGHI Nitrate Reductase A	6	7	7	C2	SLS	12910261	The oligomerization of NarGHI was analyzed using static light-scattering chromatography. The NarGHI is a dimer of a heterotrimer and is a 'flower'-shaped structure. Static light-scattering analysis of the purified NarGHI in the presence of the detergent Thesit shows a NarGHI dimer of heterotrimers. Bio interfaces are 1,2,3,4,5,6,7 but strictly transmembrane interface is only 7 (C+C chains on crystallographic 2-fold axis)
2J7A	NrfH Cytochrome C Quinol Dehydrogenase	6	2, 5, 8, 12, 14,17, 20, 23, 26, 30	23*	C2	SEC	17139260	This is a complex of NrfH (a TM chain) and NrfA (a soluble dimer): $NrfHA_2$. In crystal 2 NrfHA2 complexes (2 chains of NrfA dimer + 1 chain of NrfH) come together to form a C2 dimer super complex. The determined molecular mass of the whole NrfHA2 complex is 300 kDa by SEC, which corresponds to two $NrfHA_2$ units. There are 3 copies of the super complex in the AU. We chose as reference the copy with chains A,B,C,D,E,F: A,B,D,E are the purely soluble chains; C,F are the partly TM ones.
	Mo/Wbis-MGD Oxidoreductases							No structures from this class could be validated.
	Electron Transport Chain Complexes: Complex I							No structures from this class could be validated.
	Electron Transport Chain Complexes: Complex II							
2BS2	Fumarate Reductase Complex	6	1-8	3	C2	SEC	10586875, 17024183, 11004459	Quinol:Fumarate Reductase complex from *W. succinogenes* was solved by the authors in second ref. Fumarate Reductase was first solved by authors in first ref. The authors in third ref study the dimer formation of Fumarate Reductase (3700 A^2) by analytical gel filtration technique and found that this dimer is apparently also present in the detergent-solubilised state of the enzyme, implying that it is unlikely to be an artifact of crystallisation. Each monomeric unit is a hetero-trimer with three chains. The C2 dimer of trimers is in AU. Only chains C+F are in TM region.
	Electron Transport Chain Complexes: Complex III (Cytochrome bc1)							

1PPJ	Cytochrome bc1	10	1, 4, 6, 9, 10, 14, 16, 18, 20, 21, 24, 25, 31, 33, 36, 40	14*		SEC, EM	2982319, 6273583, 9204897, 16024040	Bovine cyt bc1. The authors in first ref did gel filtration experiments on Bovine heart cytochrome b-c1 complex. They found that the molecular conversion between the monomeric and dimeric state of the enzyme was reversible and dependent on the detergent concentration. The authors in the second ref did EM studies and found bc1 to be dimeric. Authors in third ref claim that under the conditions of activity assays the complex is most likely in the dimeric state. We take only the contacts in one monomer to be biological (10 chains: A-J). The AU contains a dimer of 2x10 subunits. As TM contacts we can take only one (with 2 copies in AU) between C+G (or P+T), it is mostly TM although there is quite some part from the smaller chain in the soluble region. All the other TM interfaces, those for chains D(Q) and E(R) with chain C(P), are much larger in the soluble region than in the TM region, so we don't take them.
3CX5	Cytochrome bc1	11	2, 3, 6, 7, 10, 17, 18, 21, 22, 24, 27, 28, 34, 36, 37, 39	6*		Homology	18390544	Yeast cyt bc1. The authors crystallize cyt bc1 complexed with cyt-c (known to be a transient interaction). The structure is similar to Bovine heart cyt bc1 complex (rmsd 1.8). Similar to previous entry we consider only monomeric units and take 9 subunits of the monomer as biological. We exclude the cyt c (chain W) which interacts only with one of the 2 monomers of the AU. Note that there's also antibody fragments in crystal (chains J(U) and K(V)). As bio interfaces we take those between the first monomer (9 chains: A-I). Interface 40 (W+O) corresponds to the transient cyt-c to cyt-bc1 transient interaction. As bio TM interfaces we take as before only the one that has interactions mostly in the TM region C+H (with another copy in AU N+S). The others have interactions mostly in soluble regions.
	Electron Transport Chain Complexes: Cytochrome b6f of Oxygenic Photosynthesis							No structures from this class could be validated.
	Electron Transport Chain Complexes: Complex IV (Cytochrome C Oxidase)							
1V54	Cytochrome C Oxidase, aa3	26	1, 4, 6, 8, 10, 14, 15, 18, 20, 21, 26, 29, 31, 33, 35, 37, 38 28, 30	4, 6, 8, 15*, 30, 31	C2	CCL, RI	8241183, Ref, 3017697	The authors in first ref did CCL experiments on Bovine Heart Mitochondrial Cytochrome c Oxidase in Triton X-100. Their results indicate that subunit I from each monomer provide one site of interaction between monomers in the dimeric form of the enzyme and that cytochrome c oxidase monomers may reassociate to form dimeric complexes in phospholipid vesicles. Radiation Inactivation studies by authors in third ref also shows that C oxidase from Bovine heart is dimeric in nature. The authors in second ref did the structure studies on Bovine Heart Mitochondrial Cytochrome and found 13 subunits in the monomer. We take as bio all interfaces of first 13-mer monomer (A to M chains) (first line) and the inter-monomer interfaces (second line), discarding anything below 400A2 As bio TM only the subset that is mostly TM (some are mostly soluble)
3HB3	Cytochrome C Oxidase, aa3	2	1	1*		SEC	3017928, 19374884	Three different detergents CsE45, dodecyl maltoside, and Triton X-t00 have been used to study the aggregation state of Paracoccus cytochrome *C* oxidase. Three different sedimentation coefficients ranging from 4.1 to 12.2 were obtained, but in all cases the enzyme proved to be monomeric. It seems thus that, for this enzyme, the monomer (of 2 subunits) is the stable and most active form, while in the mammalian oxidase the dimer is considered to be most active.
	Nitric Oxide Reductases							No structures from this class could be validated.
	Photosystems							
1JB0	Photosystem I	36	1-6, 7, 8-20	1, 5, 7	C3	SEC, EM	Ref, 11418848	The trimeric nature of the protein was predicted by the authors in first ref. The trimeric nature was proved by HPLC gel filtration. The subunit composition was analyzed by SDS gel electrophoresis. Their PSI contained 11 subunits in a monomer. The structure described in the pdb here contains 12 subunits (an extra 35 residues X subunit), but the authors do not do any experiments to confirm that the 12th subunit also sticks together. They even call it "controversial" subunit. As bio we take all interfaces in AU plus interface 7 which is the one that mediates the trimerization As bio TM we take the ones that have a significant interaction at the TM region and are not mainly mediated by soluble parts or by chlorophyls/carotenoids (the structure is more of a protein+chlorophyl+carotenoids complex)
3ARC	Photosystem II	20	1, 4, 5, 7, 9, 12, 14, 15, 16, 19, 21, 22, 23, 25, 29, 31, 33, 35, 36, 37, 43, 44, 47, 48, 52, 56, 57, 59, 61, 62, 64, 67, 70, 73	1, 4, 5		Blue Native - PAGE	Ref, 21499260, 11217865, 12518057	PS II from *Thermosynechococcus vulcanus* is very similar to those from *Thermosynechococcus elongatus*. PS II in these organisms appears as a dimer of 20 subunits in the crystal. The authors in the third ref do SDS-PAGE analysis on PSII, and all the subunits break down in SDS. The authors in first ref did a BN-PAGE analysis of PS II. They found that PS II exists in equilibrium between dimer and monomer depending on the concentration of the detergent. But all 20 subunits moved together, which indicates that 20 subunits have bio contacts, while it is not so clear what to say about the dimer As bio interfaces we take all in AU of first subunit (19 chains with capital letters) As bio TM we take the largest mostly TM ones without taking any of the peripheral helices which are mostly mediated by chlorophyls/carotenoids
	Light-Harvesting Complexes							
1LGH	Light-Harvesting Complex	16	1, 3, 5, 8	1, 3, 5, 8	C8	AUC	Ref, 8736556	The structure of LH2 from Rhodospirillum molischianum is also similar to Rps. acidophila (see above), but is an octamer of heterodimers. Here also the alpha and beta apoproteins are present as homodimers. The oligomeric studies of the complex by

								authors in first ref. was done by sedimentation equilibrium experiments in the analytical ultracentrifuge. The relative molar mass of the protein/pigment complex was found to be 114 500 ± 7% (maximum error), corresponding to the (8.0 ± 0.6)-fold of that of the basic unit. The complex is thus the octamer of the basic unit (a heterodimer). The SDS-PAGE analysis and electron micrographs were studied by authors in second ref. We choose chains A,B,D,E from AU as reference for building the 16-mer. Another 16-mer can be built from chains G,H,J,K.
1RWT	Light-Harvesting Complex LHC-II, Spinach Photosystem II	3	2, 3, 5	2, 3, 5	C3	SDS-PAGE, Homology	1885603, 15029188, 15719016	The authors in first ref characterized LHC II in barley. The protein was found to be trimeric by SDS-PAGE analysis. In higher plants the LHC II sequence is very conserved: it has around 90% sequence similarity in spinach, pea and barley. LHC II in Barley is well characterized as trimer, so the proteins here which are structurally similar to barley should also be trimers. In 1rwt there are 3 copies of the trimer in AU, thus interfaces 1-9 are bio, we take anyway chains B,G,F as reference and use only the trimer formed by interfaces 2,3,5 among them. In 2bhw there's only 1 copy in AU, thus bio interfaces are 1-3.
2BHW	Light-Harvesting Complex LHC-II, Pea Photosystem II	3	1, 2, 3	1, 2, 3	C3			
	Photosynthetic Reaction Centers							
2WJN	Photosynthetic Reaction Center	4	1, 2, 3, 4, 5	1		Homology	19743880, 2819866, Ref, 22439175, Ref, 10727607	The RC of *Blastochloris viridis* (2wjn) consists of four polypeptide subunits. The L and M subunits have five transmembrane helices each, and form a membrane-spanning heterodimer. And additional subunit (H) also participates with a single helix in the TM spanning region. The 4th subunit is C, which is out of the TM region. This protein has been studied extensively (20 structures in pdb). The RC from RB. SPHAEROIDES contains three chains (L, M and H) which are structurally very similar to that of viridis, the extra C chain is missing here. Bactereochlorophyll molecules are present between the two chains in the center (L, M). It has been characterised extensively since the 60s-70s. Ref 6 (10727607) characterises it by AUC with a weight of 100KDa which corresponds to the observed heterotrimer in AU. The RC from THERMOCHROMATIUM TEPIDUM is also structurally very similar to the viridis structure (almost identical). It also has four subunits. Note that for the 3 cases we take as TM interface only 1 (L+M), interface 2 (H+M) does have a helix in the TM region but most of the interface happens in the soluble region.
2J8C	Photosynthetic Reaction Center	3	1, 2, 3	1		AUC		
1EYS	Photosynthetic Reaction Center	4	1, 2, 3, 4, 5	1		Homology		

TRANSMEMBRANE PROTEINS: BETA-BARREL

pdb	name	size	bio interfaces	bio TM interfaces	PG	evidence	reference	comments
	Beta-Barrel Membrane Proteins: Porins and Relatives							
2FGR	Omp32 anion-selective porin	3	1	1	C3	DLS	16434398, 9761864	Homotrimer by light scattering
2ZFG	OmpF Porin	3	1	1	C3	CCL	3013869, 18636093	
2J1N	OmpC Osmoporin	3	1,2,3	1,2,3	C3	SDS-PAGE, CL	16949612, 374117	Runs as trimer in SDS-PAGE, becomes monomer upon heating
2F1C	OmpG **monomeric** porin	1	-	-		SEC	16797588, 12899633	SEC and pore gating studies: opens and closes in one step
2MPR	LamB Maltoporin	3	1,2,3	1,2,3	C3	AUC	9102468, 21640073	*S. typhimurium* sequence. AUC data (not shown) carried out on E. coli sequence which is about 80% identical (see below). Well known to be a trimer, e.g. by EM studies
1AF6	LamB Maltoporin	3	1,2,3	1,2,3	C3	AUC	9299337, 21640073	*E. coli* sequence.
1A0T	ScrY sucrose-specific porin	3	1,2,3	1,2,3	C3	SDS-PAGE	9437428	Runs as oligomer (trimer) in SDS-PAGE, becomes monomer upon heating
1UUN	MspA mycobacterial porin	8	1,2	1*,2*	C8	SDS-PAGE	14976314, 12767342	Run in SDS-PAGE with an apparent MM of ~ 100 KDa (the monomer would be 20 KDa), this species is octamer as seen in X-ray structure.
2O4V	OprP phosphate-specific transporter	3	1,2,3	1,2,3	C3	SDS-PAGE	17187075, 2834340	Runs as trimer in SDS-PAGE, becomes monomer upon heating
3VZT	PorB outer membrane protein, native structure	3	1	1	C3	SEC	8616894	SEC measurement for both natively purified and refolded PorB. Originally we had 3a2r, but it was obsoleted and replaced by 3vzt
	Outer Membrane Carboxylate Channels (Occ)						22272184	The improved crystal structures of Opd proteins were done by the authors in the paper. They renamed the proteins as Occ as they found that that OprD channels require a carboxyl group in the substrate for efficient transport. They claim that "all channels crystallizes as monomers". But they do not give any experimental evidence for the same. Some study was done on the oligomeric state of these channels in their previous versions. The results below are on their previous research.
3JTY	BenF-like Porin (putative) benzoate channel	1	-	-		SEC-MALLS	20737437	Size Exclusion Chromatography with Multi Angle Laser Light Scattering (SEC-MALLS) were done by which authors found PflBenF to be a monomer associated with one LDAO micelle.
	Beta-Barrel Membrane Proteins: Monomeric/ Dimeric							
1EK9	TolC outer membrane protein	3	1, 2, 3	1*, 2*, 3*	C3	SEC, SDS-PAGE	10879525, 9044294	The study of the oligomeric state was done in the second ref. "Purified TolC subjected to gel filtration exhibited a mass of 160–

								180 kDa corresponding to a trimer of a 51.5 kDa protein plus a micelle of betaOG. When TolC from the gel filtration was mixed with SDS sample buffer without boiling and analysed by SDS–PAGE, it gave a band much greater than 100 kDa. This was replaced by a band corresponding to the 51.5 kDa TolC monomer when samples were analysed with 8 M urea included in the SDS sample buffer"
1YC9	VceC outer membrane protein	3	1	1*	C3	Geometry, homology	15684414	The authors mention that VceC is trimeric. They give the structural similarity with the ToIC (above) as the reason. But they do not do any experiments to confirm it. But VceC shares the same overall fold as TolC and OprM, and the three structures can be superimposed with Cα rmsd below 2.0 Å, despite the very low degree of sequence identity (8.3%).
3D5K	OprM drug discharge outer membrane protein	3	1, 2, 3	1*, 2*, 3*	C3	Geometry, homology	20399187, 15507433	The crystal structure contains a C3 homo-trimer with huge interface areas, unlikely to be a crystal artifact. Structural comparison of OprM with TolC shows similar fold (Cα rmsd of 1.6 Å) in spite of a relatively low sequence identity (19%).
3PIK	CusC heavy metal discharge outer membrane protein	3	1	1*	C3	SEC	21249122	The crystal structure is similar to TolC with a C3 trimer. They perform gel filtration (no data shown) and found it to be a trimer.
2HDI	Colicin I receptor Cir in complex with Colicin Ia binding domain	1	-	-		AUC-SE	17464289	The authors used analytical ultracentrifugation to characterize wild-type Cir, colicin Ia, and complexes of the wild-type proteins, and both wild-type/mutant complexes. Cir and colicin Ia were observed to be monodisperse monomers, and they formed a 1:1 stoichiometric complex with high affinity. Note that this monomeric membrane protein in this crystal is in complex with another protein (colicin). Thus the first interfaces are actually from the membrane protein (chain A) + colicin (chain B). Interface 4 is the first A+A xtal interface (with only 295A2)
1QJP	OmpA	1	-	-		SEC, DLS	1370823, 10764596	SEC was done by the authors in first paper finding a monomer and second paper confirms it also by dynamic light scattering.
1QJ8	OmpX	1	-	-		Homology	10545325, 1987115	OmpX was crystallized by the authors in first reference. The model was found to be very similar to OmpA which is well characterized to be monomer. The rmsd of these two structures is also very low (~2A). The authors in the second ref characterized Ompx. They did SDS-PAGE analysis and found the molecular weight to be 18 kDa for a sequence of length 173. This might correspond to a monomer. None of the crystal contacts is parallel to membrane normal. Also Ail homolog below is very conserved and seems to be monomeric too. Most likely a monomer.
3QRA	Ail adhesion protein	1	-	-		SEC, Homology	22078566	The authors determined the molecular weight of Ail by using gel filtration chromatography. The unheated Ail sample runs primarily as a 14 kDa, for a sequence of length 157. This corresponds to a monomer of Ail. Also, Ail exhibits significant structural similarity to OmpX, with an rmsd of 1.7 Å. There is substantial sequence similarity between Ail and OmpX (the two proteins are 41% identical). Ompx is found as a monomer, which adds more evidence for the oligomeric state of Ail being a monomer.
1QD6	OmpLA (PldA) outer membrane phospholipase A monomer	2	1	1	C2	AUC, CCL	9013551, 11371166	OmpLA exists as monomer in its dormant state. In the presence of substrate and cofactor (Ca ions) they form a dimer. The dimer in presence of substrate and calcium ions was confirmed by the authors in first ref. They carried out analytical centrifugation and chemical cross linking experiments to confirm the dimeric state of OmpLA. In the inactive form OmpLA exits in an equilibrium between monomer and dimer. The original entry in the table was with pdb entry 1QD5, which correspond to the monomer form. 1QD6 corresponds to the dimer form.
1P4T	NspA surface protein	1	-	-		AUC, CCL	12716881	The authors performed analytical ultracentrifugation experiments as well as chemical cross-linking. Both methods (data not shown) indicated that NspA is a monomer in a detergent-containing solution.
2WJR	NanC Porin, model for KdgM porin family	1	-	-		SEC, CCL	19796645, 11773048	NanC belongs to the family of small monomeric KdgM-related porins. The NanC was found to be monomeric by authors of the first ref in gel filtration experiments. The studies on KdgM proteins was done by authors in second ref. They found that in the case of KdgM, migration in SDS-PAGE was the same whether the samples were boiled or not. Furthermore, formaldehyde cross-linking did not show any multimerization of KdgM. So NanC seems to be a monomer. The same as with any of the KdgM family members (seems to be a well established fact)
3FID	LpxR lipid A deacylase	1	-	-		SEC	19174515	Size-exclusion chromatography showed that LpxR is most likely a monomer in solution, also based on a micelle mass for DDM.
2FCP	FhuA, Ferrichrome-iron receptor without ligand	1	-	-		SEC, AUC	9856937, 9865695, 8916906	In both studies they found a monomer in the crystal. But they do not do any further studies to confirm it. The authors in first reference say that "In contrast to the typical trimeric arrangement found in porins, FhuA is monomeric". In third reference they characterised it as monomer by SEC and AUC.
	Outer Membrane Autotransporters						15866036	There are two known subtypes of autotransporters: monomeric autotransporters (also referred to as classical/conventional autotransporters) and trimeric autotransporters.
1UYN	NalP autotransporter translocator domain	1	-	-		SEC-MALS, CCL	21306302	They did SEC-MALS: "SEC-MALS demonstrated that BrkAβ exists as a monomer under different detergent conditions, even for the detergent C_8E_4 that was used for crystallization" Also the structures of NalP is similar to Hia (rmsd 2.6A using alignment and COOT) and EspP (full structure rmsd 3.3) (see below), providing further evidence
2GR8	Hia1022-1098 trimeric autotransporter	3	2, 3, 4	2, 3, 4	C3	SDS-PAGE	16688217	Hia was purified using a streptavidin-affinity column and then ion exchange chromatography, in each case yielding a single band on an SDS–PAGE gel corresponding to the molecular weight of a trimer. The protein is trimeric in solution.
3SLJ	EspP autotransporter, post-cleavage state	1	-	-		Blue native PAGE, AUC	16262782, 22094314	Second ref corresponds to the crystal structure of EspP. The authors in first ref did Blue Native polyacrylamide gel electrophoresis, analytical ultracentrifugation and other biochemical methods showing that EspP behaves as a compact monomer.
3AEH	Hbp (hemoglobin protease) self-cleaving autotransporter with truncated passenger	1	-	-		Homology	20615416, 15728184	The authors compare the structure to NalP and EspP autotransporters. It is very similar to EspP (rmsd 0.4). The authors also did SEC for purification (data not shown) but nothing is mentioned about the oligomeric state based on SEC.

3KVN	EstA Autotransporter, full length	1	-	-		Homology	20060837	The authors here give a full structure of the autotransporter, but the basic structure is similar to those of EspP and NalP that are well characterized, so EstA must also exist as monomer.
3ML3	IcsA autotransporter (autochaperone region only)	1	-	-		SEC	21335457	In the crystal lattice, IcsA-AC is arranged as a head-on-head dimer. In solution, the protein appears to be monomeric. IcsA-AC elutes in a single peak from a gel filtration column. The molecular mass calculated from the elution volume is 26 kDa, and the expected molecular mass of a monomer is 20 kDa.
	Omp85-TpsB Outer Membrane Transporter Superfamily							No structures from this class could be validated.
	Beta-Barrel Membrane Proteins: Mitochondrial Outer Membrane							No structures from this class could be validated.
	Adventitious Membrane Proteins: Beta-sheet Pore-forming Toxins							
7AHL	alpha-hemolysin	7	1, 2, 3, 4, 5, 6, 7	1*, 2*, 3*, 4*, 5*, 6*, 7*	C7	SDS-PAGE	8943190, 21280135, 6272304, 9512705	The structure was solved by the authors in first two ref. They get a heptamer in the crystal structure. SDS-PAGE analysis was done by the authors in third ref. They find the molecular weight to be of a hexamer, but there might be errors in the calculation which can easily lead to a heptamer. AFM studies were done by authors in last ref. AFM images clearly show a hexamer. In any case the extensive interfaces in crystal and amazing symmetry are unlikely to be artifacts. Note that all interfaces have a TM part and a soluble part.

4 Conclusions

This thesis offers a comprehensive study on protein-protein interfaces, a topic central to structural biology. These interfaces, being mediators of protein-protein interactions, are in effect responsible for much of the complexity of biological systems.

Many aspects of protein interfaces were investigated. By studying them we mainly aimed at solving an important problem inherent to protein X-ray crystallography: the classification of crystal interfaces into biologically relevant ones (specific) and crystal lattice contacts (non-specific). Both soluble and transmembrane protein interfaces were studied, contributing to the generality of the results. The most important outcomes of this work are summarised below.

A new effective method for the computational classification of crystal interfaces was proposed and shown to be accurate by using sets of validated biological and crystal contacts. The method relies mostly on evolutionary data to classify interfaces, thus offering a completely new perspective compared to the alternative available methods for solving the problem. Furthermore the classification was tested with good success in both soluble and transmembrane protein interfaces.

New datasets of validated biological and crystal interfaces were compiled. These datasets constitute a valuable resource in their own as they represent the first available datasets that center on the difficult to classify interface area region. Thanks to them important aspects of the interfaces could be uncovered, especially the fact

that the presence of fully buried residues –interface core residues– is a necessary condition for an interface to be specific.

Additionally a new dataset –the first of its kind– of validated transmembrane protein-protein interfaces was compiled offering a test-base for the analysis of transmembrane interfaces. Among other things it allowed us to establish the validity of the principles observed for soluble proteins also in membrane protein interfaces.

Using conservative sequence alignments and simple sequence conservation measures allowed us to uncover the features present on protein surfaces and in particular to detect the footprint of evolution on protein interfaces. The robustness of the predictions and their future potential were demonstrated in a retrospective analysis that used historical sequence data of the last 10 years.

The applicability of the method to important structural biology problems was shown by the study of putative dimerization interfaces of G-protein coupled receptors (GPCRs). That analysis clearly showed the potential of EPPIC in offering new hypotheses and as a complement that enhances the structural data.

As a final but very important outcome, a robust implementation of the method was developed, both as a command-line software package and as a web user interface, made available to the structural biology and bioinformatics communities through a web server. The availability of such tool should contribute to discoveries and new hypotheses in structural biology.

4.1 Applications to Structural Biology

As a demonstration of the predictive power of our EPPIC evolutionary method we now present a few prominent examples of crystal structures where our method not only clarifies the possible biological interfaces present in the crystal but also offers new biological hypotheses based on the structural data.

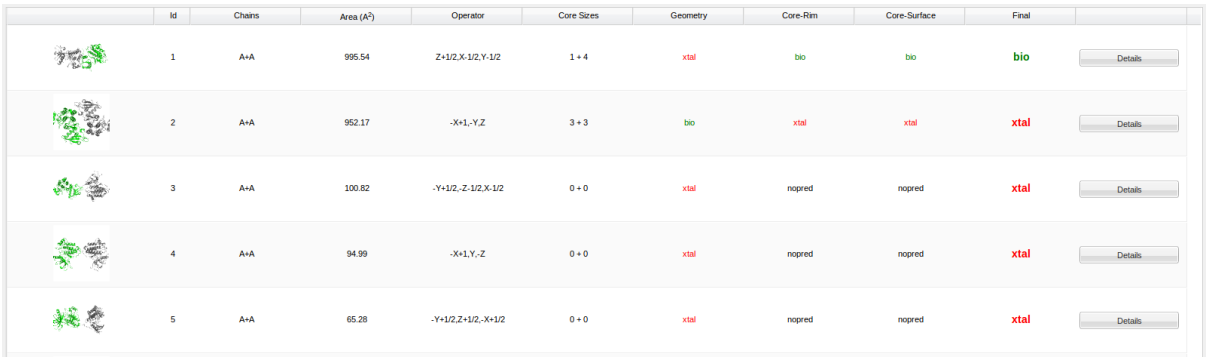
4.1.1 The kinase domain of epidermial growth factor receptor

The structure of the EGFR kinase domain was solved by Zhang et al in 2006 [1] [PDB: 2gs2]. They proposed that the activation mechanism of the EGFR is based on the dimerization they observed in the structure. Two possible putative dimer interfaces are present in the crystal structure: a symmetric one on a 2-fold axis with an area of ~950 Å^2 and an asymmetric one, with a slightly larger interface area.

EPPIC shows a very clear biological signal for the asymmetric interface while the symmetric interface is classified as crystal contact by the two evolutionary indicators. An asymmetric (or heterologous in Monod's nomenclature, see the Introduction to this thesis) interface is in principle not viable since it leads to infinite

helical-like assemblies. However in this context, the protein solved constitutes only one of the (soluble) domains of a much larger transmembrane protein and thus such heterologous interface is viable in the full length assembly.

The prediction of EPPIC confirms the dimerization mechanism that Zhang et al proposed, validated by them through a series of mutagenesis experiments. They provided further support to the hypothesis by pointing out that the two EGFR kinase domains interact analogously to CDK2 and cyclinA in their complex [2], with EGFR monomer A (the activated kinase) corresponding to CDK2 and monomer B corresponding to CyclinA.



Id	Chains	Area (A^2)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	A+A	995.54	Z+1/2,X-1/2,Y-1/2	1 + 4	xtal	bio	bio	bio	Details
2	A+A	952.17	-X+1,-Y,Z	3 + 3	bio	xtal	xtal	xtal	Details
3	A+A	100.82	-Y+1/2,-Z-1/2,X-1/2	0 + 0	xtal	nopred	nopred	xtal	Details
4	A+A	94.99	-X+1,Y,-Z	0 + 0	xtal	nopred	nopred	xtal	Details
5	A+A	65.28	-Y+1/2,Z+1/2,-X+1/2	0 + 0	xtal	nopred	nopred	xtal	Details

Figure 1 A screenshot of the EPPIC web server with the predictions for the first interfaces of 2gs2. Interface 1 is the one identified in the study of Zhang et al as the relevant one, whilst interface 2 is the symmetric one that they see as not relevant. The prediction comes in 3 separate columns for each of the 3 indicators plus a "Final" column with the final decision. The first column is for the geometrical indicator while the other two columns are the two evolutionary indicators: core-rim score and core-surface score.

4.1.2 Human RhoA and the effector domain of the protein Kinase PKN/PRK1

The structure of human RhoA complexed with the effector domain of the protein kinase PRK1 (also referred to as PKN) was solved in 1999 by Maesaki et al [3] [PDB: 1cxz]. The crystal contains two fairly large interfaces between the RhoA and PRK1, referred to by the authors as "contact 1" and "contact 2". The chosen asymmetric unit is the one containing contact 1 (largest interface) and it is the interface towards which most of the analysis in the paper is focused. The authors do acknowledge in any case that contact 2 is a potentially valid interface and end up concluding that both interfaces are relevant. They based the hypothesis on existing evidence of RhoA binding PRK1 in a 2:1 stoichiometry [4], although in the crystal they can only measure a 1:1 stoichiometry from SDS-PAGE and time-of-flight type MS [3].

Later Modha et al [5] solved by NMR the structure of the homologous protein Rac1 complexed with its effector PRK1 [PDB: 2rmk]. The interface observed in that NMR complex was analogous to contact 2 of the earlier crystal structure.

Finally in 2011 the same group settled the issue with a mutagenesis analysis [6] where they clearly demonstrate that "contact 2" is the only valid interaction interface

for the RhoA and PRK1 complex. In an Alanine scanning experiment they select residues to mutate from both sites (contact 1 and contact 2), using a scintillation proximity assay to measure an apparent K_d. The residues that significantly reduced the affinity in the Alanine scan were all located in contact 2.

We analysed with our EPPIC method the original crystal structure [PDB: 1cxz] and found a clear crystal signal for interface 1 and a very clear biological signal for interface 2. Despite its smaller size in terms of buried area, interface 2 also contains many more core residues than interface 1. The evolutionary analysis of both core-rim and core-surface indicators also confirms the geometry indication based on alignments of 105 homologs for the RhoA (chain A) and 15 homologs for the PRK1 (chain B).

Figure 2 The EPPIC output screenshot for 1cxz. The biological signal for interface 2 is very clear from all 3 indicators, whilst interface 1 is predicted to be a crystal contact.

This particular case thus offers a fantastic example of how EPPIC can be applied to specific problems in structural biology and enhance the structural data with the predictive power of evolutionary signal provided by the sequence data.

4.1.3 The Zinc transporter CzrB from Thermus Thermophilus

Zinc transporters are a large family of transmembrane proteins with representatives in eukaryotes like zinc tranporter-3 (ZnT-3) and ZnT-8 and also in bacteria, with representatives like YiiP in *E Coli* or CzrB in *Thermus Thermophilus*. The cytoplasmic C-terminal domain of CzrB was solved by Cherezov et al [7] in both the Apo [PDB: 3byp] and Zn-bound [PDB: 3byr] forms. The Apo (Apo-CzrB) and Zn-bound (Zn-CzrB) structures are almost identical in terms of tertiary structure, but in terms of their interfaces there is an important difference in a putative dimer interface seen in both, being the interface area for Zn-CzrB (~950 Å²) much larger than that of Apo-CzrB (~400 Å²). The authors propose that both structures are dimers, having also some additional data from Size Exclusion Chromatography, SAXS and NMR to support it. From there they hypothesize a mode of action for the transporter based on modelling done with the homologous full length structure of YiiP from *E Coli* [PDB: 2qfi] published in 2007 [8]. A higher resolution structure of YiiP [PDB: 3h90]

came only a couple of years later [9]. The sequence identity between the two full length proteins is around 32%, but structurally they seem to be extremely well conserved. A superposition of YiiP and the C-terminal domain of Zn-CzrB produces an almost perfect fit not only for the tertiary structure but also for the dimer interface.

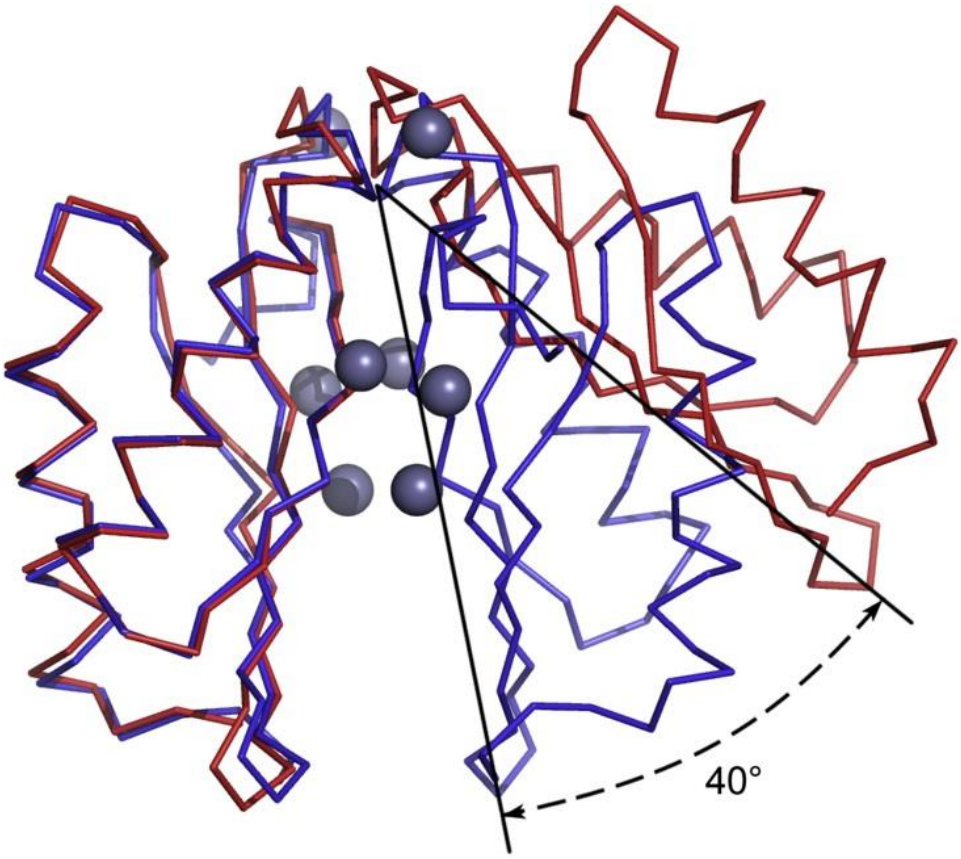


Figure 3 The proposed hinge of Cherezov et al: the blue ribbon corresponds to the interface seen in the Zn-bound structure (3byr) whilst the red one is the Apo structure (3byp). Adapted from Figure 2 in Cherezov et al [7]

Cherezov et al proposed a possible mode of action based on their Apo and Zn-bound structures: the full length transporter goes from an open (Apo) to closed (Zn-bound) conformation. Only the closed one allows enough space for a chaperone to bind laterally.

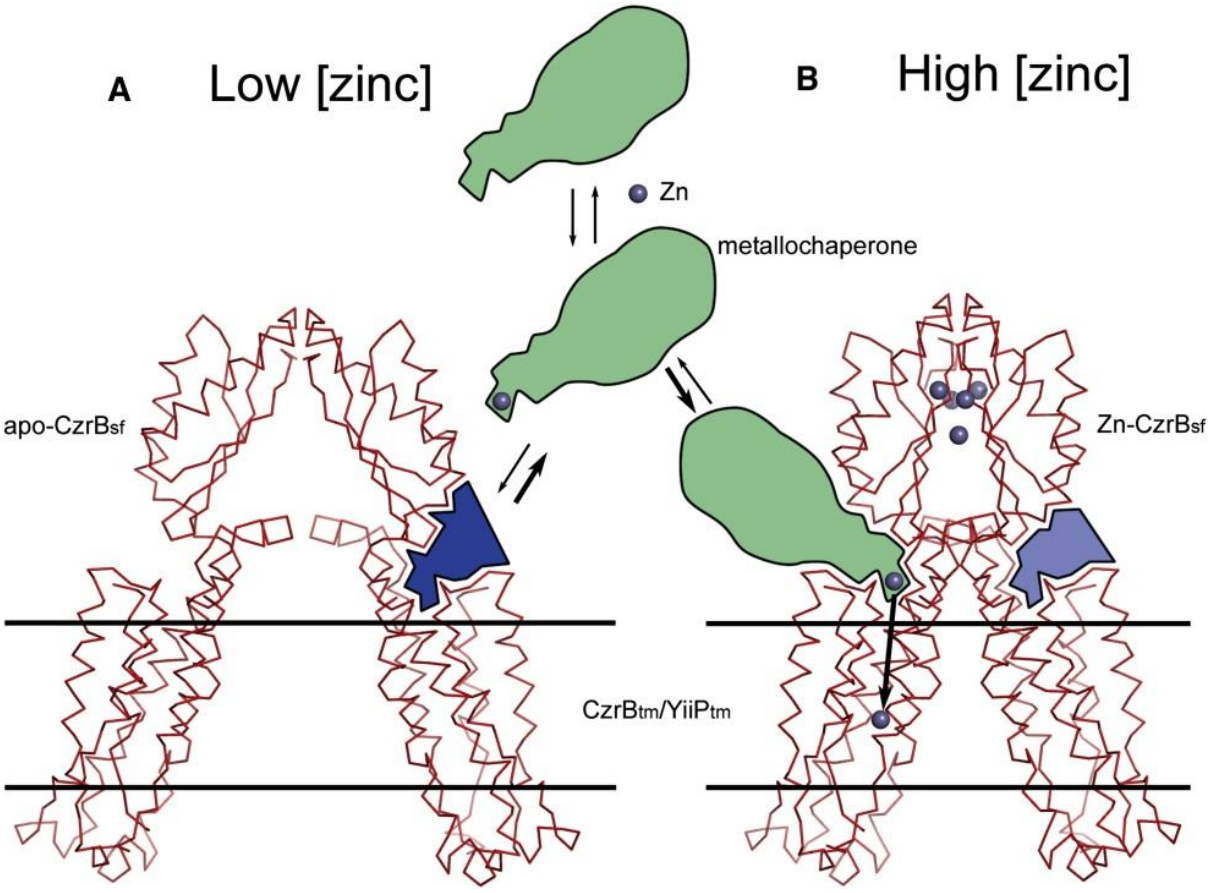
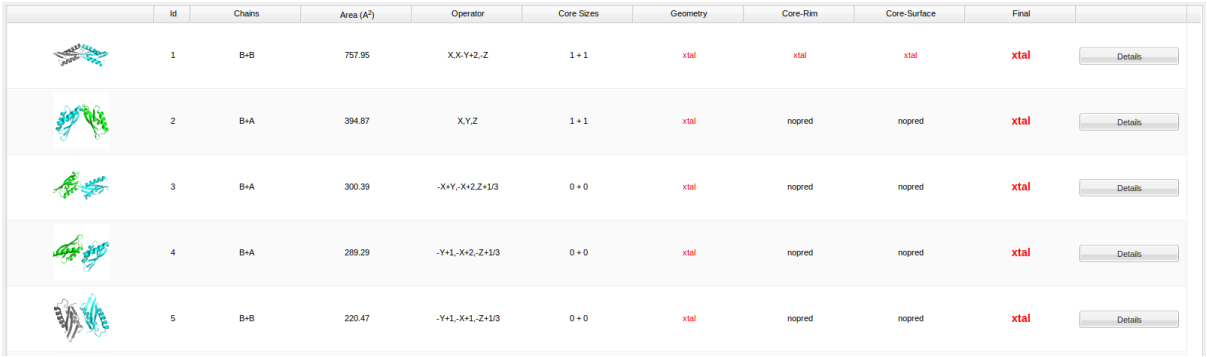


Figure 4 The proposed mode of action of the full length CzrB Zn transporter. Left the Apo structure showing the putative "open" interface and right the Zn-bound structure showing the "closed" interface. Adapted from Figure 8 in Cherezov et al [7]

However, looking at EPPIC classification for interfaces of both structures (3byp and 3byr) it is striking to see that the Zn-bound form (3byr) presents a clear biological signal from the evolutionary indicators, whilst in the Apo structure (3byp) the proposed dimerization interface has too little packing to even measure a biological signal on it. In fact its area falls into what the EPPIC classifier calls the hard area limit for crystal contact [10]. Thus following these predictions, we would rather propose that the Apo dimer interface is just a crystal contact and does not say much about the unbound state. Experimental evidence was given by the authors for the Apo molecule to behave as dimer in solution. We can only hypothesize that some kind of weak dimer equilibrium exists, but based instead on the 1st interface seen in the crystal (with an area of ~750 Å^2).

Id	Chains	Area (Å²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	A+A	957.47	X,-Y+1,-Z+1	3 + 3	bio	bio	bio	bio	Details
2	A+A	513.14	-X+1,Y,-Z+1/2	0 + 0	xtal	nopred	nopred	xtal	Details
3	A+A	484.25	-X+1,Y,-Z+3/2	1 + 1	xtal	xtal	xtal	xtal	Details
4	A+A	166.48	X-1/2,-Y+1/2,-Z+1	0 + 0	xtal	nopred	nopred	xtal	Details
5	A+A	149.02	-X,Y,-Z+1/2	0 + 0	xtal	nopred	nopred	xtal	Details

Figure 5 The EPPIC output screenshot for 3byr showing the clear biological signal foun in interface 1



Id	Chains	Area (Å²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	B+B	757.95	X,X-Y+2,-Z	1 + 1	xtal	xtal	xtal	xtal	Details
2	B+A	394.87	X,Y,Z	1 + 1	xtal	nopred	nopred	xtal	Details
3	B+A	300.39	-X+Y,-X+2,Z+1/3	0 + 0	xtal	nopred	nopred	xtal	Details
4	B+A	289.29	-Y+1,-X+2,-Z+1/3	0 + 0	xtal	nopred	nopred	xtal	Details
5	B+B	220.47	-Y+1,-X+1,-Z+1/3	0 + 0	xtal	nopred	nopred	xtal	Details

Figure 6 The EPPIC output screenshot of 3byp. Interface 1 was proposed as relevant but no signal can be seen by the EPPIC software. Additionally the interface area is extremely small for what is typical in biological interfaces.

Our hypothesis is backed by the high resolution Zn-bound full-length structure of the *E coli* homolog YiiP [PDB: 3h90]. It dimerizes not only through the interface in the C-terminal domain seen above but also through an important "charge interlock" salt bridge occurring at the C-terminal end of the transmembrane domain [9], an interaction that would not be possible to satisfy in the above proposed open conformation. Further the models based on EM maps of the Apo structure [11] also do not support the mode of action proposed by Cherezov and colleagues. The 2010 review by Dax Fu [12] also puts the model into doubt by stating "the lack of a critical dimeric association in the CzrB fragment structures raises the question as to the functional relevance of the observed conformational change".

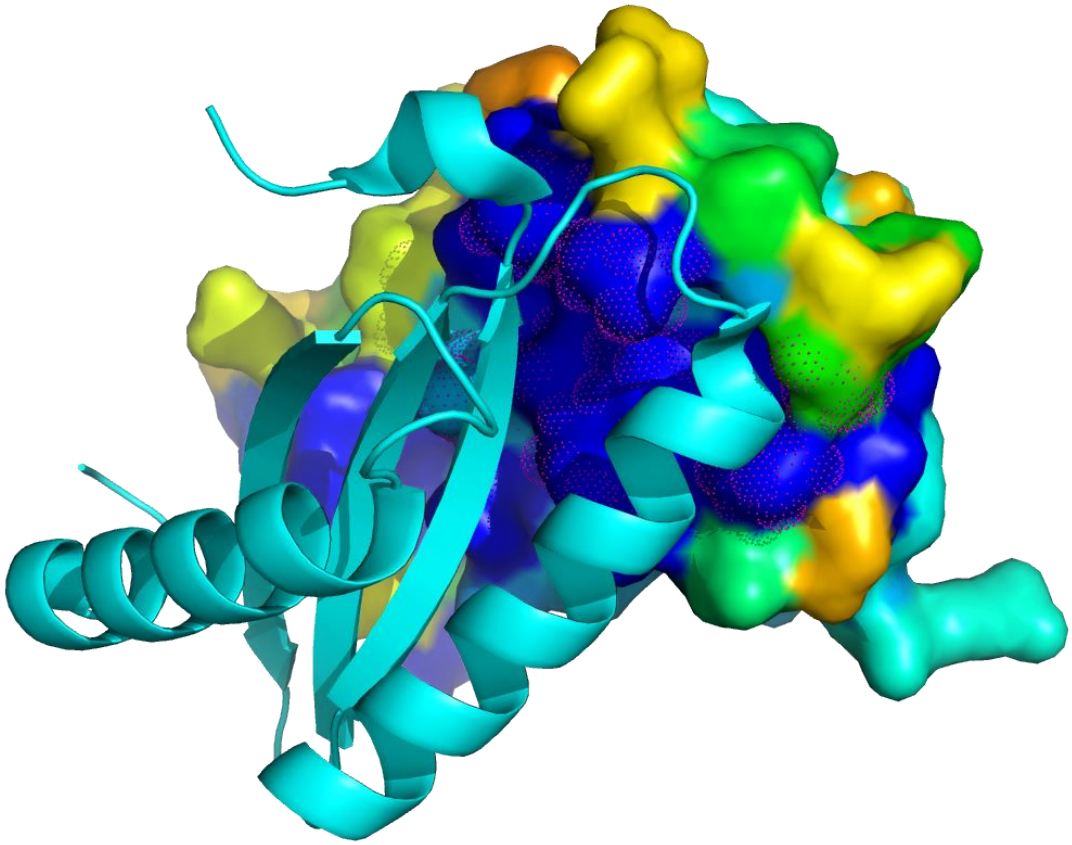


Figure 7 The interface 1 of the Zn-bound CzrB transporter [PDB: 3byr]. One of the chains is depicted with a surface representation where coloring indicates sequence entropy values (high conservation in blue, going towards yellow and orange in lower conservation). The other chain is depicted in cyan in cartoon representation. The residues that constitute the core of the interface (burial >70%) are shown with magenta dots. The interface is very clearly located at the well conserved patch.

4.2 Shortcomings

We have been able to identify a few issues in the current implementation of the EPPIC method which represent weak points and lead in some cases to interface misclassifications or wrong assembly inference.

An example of such an issue which we were able to partially correct was that of the **treatment of large ligands**. We came across the problem by studying haemoglobin, which in most vertebrates is a well-established tetramer composed of 2 alpha and 2 beta subunits in a D2 pseudosymmetry. The protein has been solved many times and thus has many representatives in the PDB, for human and also other vertebrate species. In many of the structures analysed the two core-rim and core-surface evolutionary indicators had high values above the crystal contact cut-off for the two biological interfaces of the assembly, thus resulting in the wrong classification of the

interfaces. By looking at the sequence entropies color-mapped onto the protein surface, we realized that the main reason behind it was the heme pocket: in haemoglobin very deep pockets in each of the protomers hold the heme molecules and residues surrounding the heme are extremely well conserved.

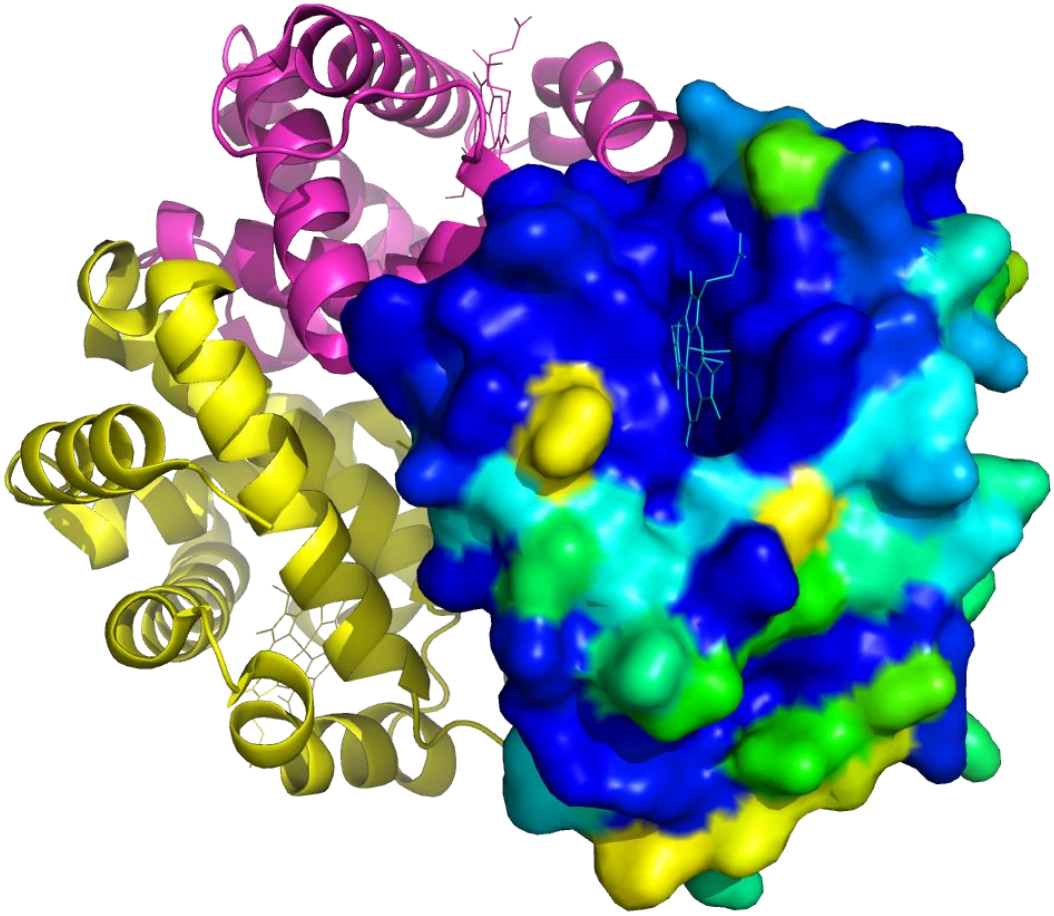
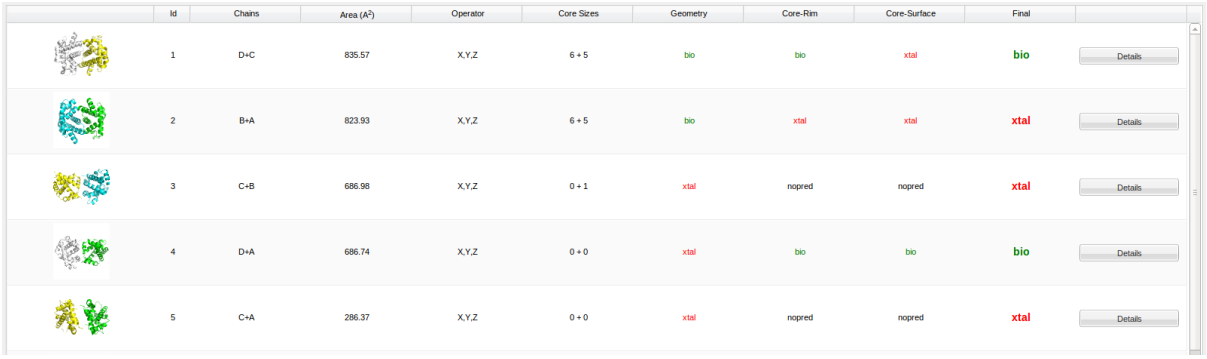


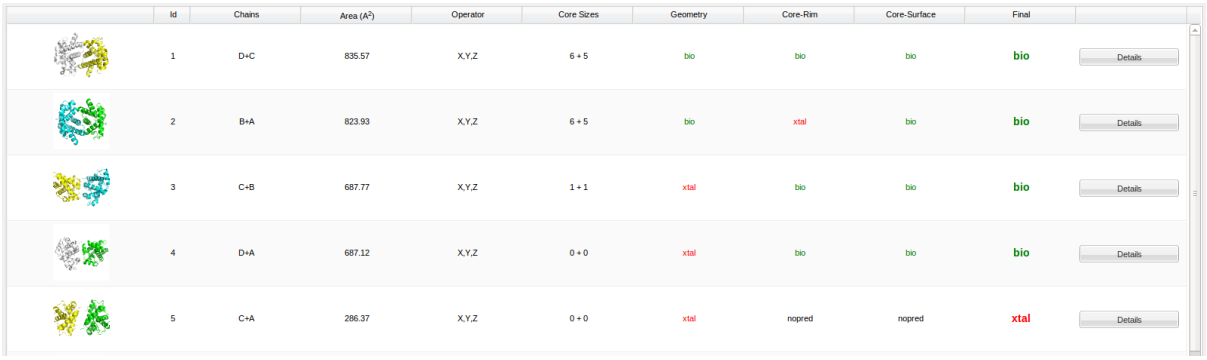
Figure 8 A representation of the hemoglobin tetramer [PDB: 2d5z]. One of the alpha subunits is shown in surface representation colored by sequence entropy values. The other chains are in cartoon representations. The heme molecule can be seen in its deep pocket where residues lining it are very well conserved.

Those residues were then considered as surface for the core-surface indicator and thus were biasing the background entropy distribution towards lower (more conserved) values. The solution that we decided to implement was that of treating large ligand molecules as attached to their corresponding protomers in order to calculate ASAs. In that way the residues in direct contact with the cofactor would not then be counted as surface residues but as protein interior. This procedure could solve the problem for at least some of the haemoglobin cases, see for instance the structures [PDB: 2d5z] or [PDB: 2dn3].



Id	Chains	Area (Å²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	D+C	835.57	X,Y,Z	6 + 5	bio	bio	xtal	bio	Details
2	B+A	823.93	X,Y,Z	6 + 5	bio	xtal	xtal	xtal	Details
3	C+B	686.98	X,Y,Z	0 + 1	xtal	nopred	nopred	xtal	Details
4	D+A	686.74	X,Y,Z	0 + 0	xtal	bio	bio	bio	Details
5	C+A	286.37	X,Y,Z	0 + 0	xtal	nopred	nopred	xtal	Details

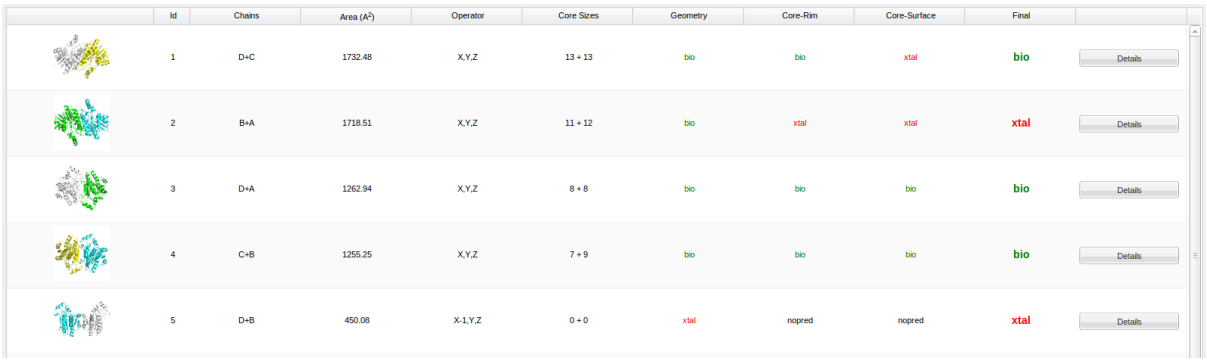
Figure 9 The EPPIC output screenshot for 2d5z. All residues in surface, including the ones facing the heme molecule, were considered for the core-surface score. The final prediction fails for interfaces 2 and 3.



Id	Chains	Area (Å²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	D+C	835.57	X,Y,Z	6 + 5	bio	bio	bio	bio	Details
2	B+A	823.93	X,Y,Z	6 + 5	bio	xtal	bio	bio	Details
3	C+B	687.77	X,Y,Z	1 + 1	xtal	bio	bio	bio	Details
4	D+A	687.12	X,Y,Z	0 + 0	xtal	bio	bio	bio	Details
5	C+A	286.37	X,Y,Z	0 + 0	xtal	nopred	nopred	xtal	Details

Figure 10 The EPPIC output screenshot for 2d5z where now heme pocket residues are not considered to be at the surface in order to calculate the core-surface score. The prediction changes from "xtal" to "bio" in interfaces 2 and 3.

Another example of a molecule where a ligand caused bias in the surface entropy distributions can be seen in the structure of the mouse lung carbonyl reductase [PDB: 1cyd] [13]. The protein is a tetramer and uses NADPH and NADH as coenzymes. As with haemoglobin the cofactors sit in deep pockets in the surfaces of each of the monomers, where residues lining the pocket are very well conserved. The EPPIC analysis by considering and not considering the ligands vary quite dramatically.



Id	Chains	Area (Å²)	Operator	Core Sizes	Geometry	Core-Rim	Core-Surface	Final	
1	D+C	1732.48	X,Y,Z	13 + 13	bio	bio	xtal	bio	Details
2	B+A	1718.51	X,Y,Z	11 + 12	bio	xtal	xtal	xtal	Details
3	D+A	1262.94	X,Y,Z	8 + 8	bio	bio	bio	bio	Details
4	C+B	1255.25	X,Y,Z	7 + 9	bio	bio	bio	bio	Details
5	D+B	450.08	X-1,Y,Z	0 + 0	xtal	nopred	nopred	xtal	Details

Figure 11 The EPPIC output screenshot for 1cyd where residues in contact with the cofactors are considered as surface in the core-surface score calculation. Interface 2 is predicted wrongly as crystal contact.

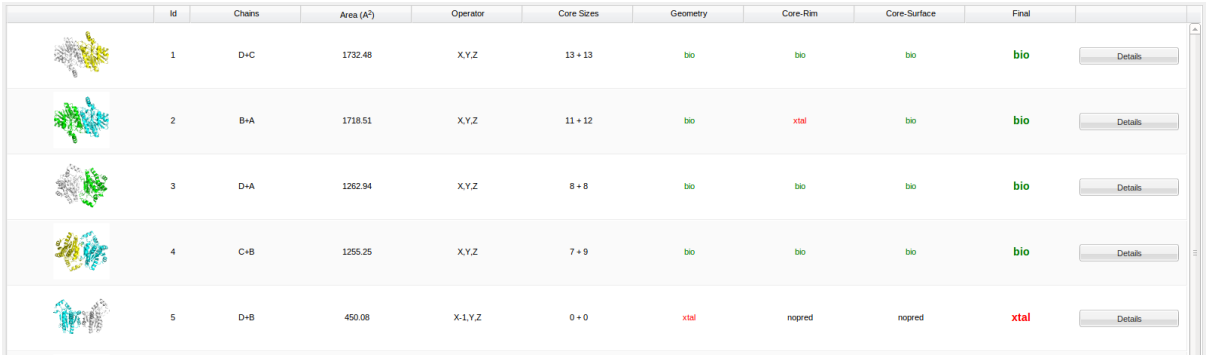


Figure 12 The EPPIC output screenshot for 1cyd, this time residues in contact with the cofactors were not considered as surface residues for the core-surface score calculation.

Another issue that has produced misclassifications is that of **poor multiple sequence alignments**. As described in the Duarte el al paper [14] we use a 60% identity cut-off for selecting homologs, going down to 50% if not enough homologs exist. The strategy clearly showed advantages when measuring the performance of the method in our datasets of biological interfaces. However a problematic issue appears when only very similar-to-query sequence homologs exist, e.g. an alignment that contains sequences with identities around 90% but no other sequences with lower identities. The entropy variability measure in such cases has a low information content, in fact most of the alignment columns will show an entropy value of 0 or very close, leading to very little detectable signal at the interfaces. An example of such a case is that of sheep's alpha-beta tubulin heterodimer [PDB: 4drx] [15]: tubulin is an extremely well conserved protein in all eukaryotes, from human to plants or even to yeast. Sequence identities for most homologs are above 80% or even 90% to that of the sheep tubulin, thus the alignment that EPPIC can produce has reduced information content, making it more problematic finding evolutionary signal at the surface.

The case where a big gap in sequence knowledge for the particular query is present is also problematic: known sequence homologs can be found only in the 50% identity-to-query region but no other homologs in between. An example would be the structure of quinate dehydrogenase from *Corynebacterium Glutamicum* [PDB: 3jyo] where almost all sequence homologs known are around the 50% identity to the query.

In practice this kind of poor alignments tends to happen most often in bacterial, archaeal or viral proteins where proteins evolve at the fastest pace. That combines with the fact that the sampling of sequence space is still very limited for the bacterial world in comparison to that of high eukaryotes. Hopefully lower costs and even higher throughputs of sequencing technologies enabling more metagenomics studies will make this problem disappear eventually.

As a final issue frequently found, we can also mention classification problems in **small proteins** or in proteins with very large interfaces compared to their free surfaces. In these cases it is not possible for the core-surface method to make a prediction as not enough residues are available to sample properly the background entropy distribution. The same problem surely applies as well to peptide-protein interfaces, where we find difficulty in assigning scores to the peptide side of the interface. Additionally the peptide case presents challenges in finding homologs when the sequences are very short. The case of the rotor ring of Na-dependent F-ATP synthase [PDB: 2wgm] [16] is an example where no core-surface score can be calculated since the monomers are small compared to the interfaces.

4.3 Outlook

The methods developed in this thesis have demonstrated to offer a lot of potential in understanding protein crystal structures in the light of evolution. These results are then very encouraging for exploring further paths and continuing development in the project.

The main missing piece in the project is most obviously the automatic inference of quaternary structures based on the crystal structure. At the moment we are able to provide an opinion of whether a particular interface, in which a pair of protein chains comes together, represents a bona-fide biologically relevant interface. From a set of all pairwise biologically relevant interfaces in a crystal one can in principle construct the full assembly most likely constituting the biological assembly. However our EPPIC method cannot at the moment assemble the different interfaces automatically. Such a development would surely facilitate many analyses in bioinformatics or structural biology.

Most importantly the implementation of such a method would aid enormously in the task of classifying pairwise interfaces. It is only in the context of the full assembly in the protein crystal where all factors can be properly assessed in order to predict a biological unit. In here a fundamental feature, ignored by the EPPIC method so far, should bring a lot of new insights: symmetry.

Symmetry is a prevalent feature of biological macromolecules and one of the most prominent "driving forces" of protein oligomeric assemblies. In fact as presented in the Introduction it is a necessity in the case of homomers. Assembly of multiple pieces of the same monomer can only occur if the surfaces that interact are all satisfied at the same time in a closed symmetry. Otherwise the protein would aggregate into infinite fiber-like assemblies [17]. But symmetry does not stop in homomers, it is also pervasive in heteromeric structures. The PDB has in fact as of May 2013 out of 84000 protein structures, 22000 C2 symmetry molecules, 2800 with

C3 symmetry, 1400 with other cyclic symmetries, 4500 D2, 1500 D3 and 857 other dihedral symmetries. A further 800 molecules present icosahedral, tetrahedral or octahedral symmetries.

We anticipate that a combination of our pairwise interface classification together with symmetry considerations will enhance greatly the predictive power of the evolutionary indicators. The most straight forward idea to implement comes from the fact that crystal contacts are unlikely to happen in high symmetries. They are surely not unusual in the case of the most simple point group symmetry C2, as already pointed out by Bahadur et al [18]. The authors compiled a dataset of what they call "crystal dimers", i.e. crystal contacts occurring at 2-fold axes, thus resulting in C2 assemblies. On the contrary evidence for crystal contacts occurring in higher point group symmetries is scarce. In fact in the case of cyclic symmetry, crystal contacts are in principle only possible in point groups C2, C3 or multiples thereof. Of course the asymmetric unit can then produce any other kind of point group symmetry. We would then hypothesize that any observed point group symmetry in the crystal with interfaces of significant areas, constitutes a very strong indication of a biological assembly, provided that the symmetry is Cn with n>2 or Dn.

Another important idea for future explorations is that of interface prediction and protein-protein docking. We have seen so far that the evolutionary signal is clear enough in order to classify interfaces given their precise locations at the protein surfaces. The next logical step is that of finding the location of the interface, given the structure of just one of the protomers. The problem presents clearly many more challenges and would require development of an algorithm capable of rapidly sampling the different potential binding patches in order to measure their evolutionary signals. Combination of commonly used methods in the docking field would most definitely be needed to boost the predictive power.

A few other new ideas are still to be developed, for instance we envisage that our method could prove very useful in validation of crystal structures or even in validation of oligomeric models created with computational methods. Exploring the potential of EPPIC in the context of validation would require extensive analysis by using redundant PDB structures with different qualities or sets of decoy models for which a crystal structure is known.

Surely more applications of the main method are still possible and can bring new insights into structural biology problems. For instance hybrid experimental methods that aim at determining the structure of super-complexes like the Nuclear Pore Complex would benefit from our evolutionary predictions. In those cases the assembly of the full complex needs to be reconstituted from the different crystal structures of the parts and built into low resolution EM maps.

All in all we are hopeful that the ideas developed in this thesis will be able to bring new knowledge, offer hypotheses to test and in general enhance the wealth of structural and sequence data available.

References

1. Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J: **An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor.** *Cell* 2006, **125**:1137–49.

2. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massagué J, Pavletich NP: **Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex.** *Nature* 1995, **376**:313–20.

3. Maesaki R, Ihara K, Shimizu T, Kuroda S, Kaibuchi K, Hakoshima T: **The structural basis of Rho effector recognition revealed by the crystal structure of human RhoA complexed with the effector domain of PKN/PRK1.** *Molecular cell* 1999, **4**:793–803.

4. Maesaki R, Shimizu T, Ihara K, Kuroda S, Kaibuchi K, Hakoshima T: **Biochemical and crystallographic characterization of a Rho effector domain of the protein serine/threonine kinase N in a complex with RhoA.** *Journal of structural biology* 1999, **126**:166–70.

5. Modha R, Campbell LJ, Nietlispach D, Buhecha HR, Owen D, Mott HR: **The Rac1 polybasic region is required for interaction with its effector PRK1.** *The Journal of biological chemistry* 2008, **283**:1492–500.

6. Hutchinson CL, Lowe PN, McLaughlin SH, Mott HR, Owen D: **Mutational analysis reveals a single binding interface between RhoA and its effector, PRK1.** *Biochemistry* 2011, **50**:2860–9.

7. Cherezov V, Höfer N, Szebenyi DME, Kolaj O, Wall JG, Gillilan R, Srinivasan V, Jaroniec CP, Caffrey M: **Insights into the mode of action of a putative zinc transporter CzrB in Thermus thermophilus.** *Structure (London, England : 1993)* 2008, **16**:1378–88.

8. Lu M, Fu D: **Structure of the zinc transporter YiiP.** *Science (New York, N.Y.)* 2007, **317**:1746–8.

9. Lu M, Chai J, Fu D: **Structural basis for autoregulation of the zinc transporter YiiP.** *Nature structural & molecular biology* 2009, **16**:1063–7.

10. Duarte JM, Srebniak A, Schärer MA, Capitani G: **Protein interface classification by evolutionary analysis.** *BMC bioinformatics* 2012, **13**:334.

11. Coudray N, Valvo S, Hu M, Lasala R, Kim C, Vink M, Zhou M, Provasi D, Filizola M, Tao J, Fang J, Penczek PA, Ubarretxena-Belandia I, Stokes DL: **Inward-facing conformation of the zinc transporter YiiP revealed by cryoelectron microscopy.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:2140–5.

12. Fu D: **Zinc Transporter YiiP from Escherichia coli**. In *Handbook of Metalloproteins*. John Wiley & Sons, Ltd; 2006.

13. Tanaka N, Nonaka T, Nakanishi M, Deyashiki Y, Hara A, Mitsui Y: **Crystal structure of the ternary complex of mouse lung carbonyl reductase at 1.8 A resolution: the structural origin of coenzyme specificity in the short-chain dehydrogenase/reductase family.** *Structure (London, England : 1993)* 1996, **4**:33–45.

14. Duarte JM, Srebniak A, Schärer M a, Capitani G: **Protein interface classification by evolutionary analysis.** *BMC bioinformatics* 2012, **13**:334.

15. Pecqueur L, Duellberg C, Dreier B, Jiang Q, Wang C, Plückthun A, Surrey T, Gigant B, Knossow M: **A designed ankyrin repeat protein selected to bind to tubulin caps the microtubule plus end.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:12011–6.

16. Meier T, Krah A, Bond PJ, Pogoryelov D, Diederichs K, Faraldo-Gómez JD: **Complete ion-coordination structure in the rotor ring of Na+-dependent F-ATP synthases.** *Journal of molecular biology* 2009, **391**:498–507.

17. Monod J, Wyman J, Changeux JP: **On the nature of allosteric transitions: a plausible model.** *Journal of molecular biology* 1965, **12**:88–118.

18. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces.** *Journal of molecular biology* 2004, **336**:943–55.

Appendix

As mentioned in the Introduction, the development of the OWL software library was instrumental for the implementation of the main project of this thesis. The development predates the work carried out at the Paul Scherrer Institute and was performed together with colleagues at Michael Lappe's group at the Max Planck Institute for Molecular Genetics in Berlin. The main topic of research there was that of protein structure prediction and analysis via the use of network representations. In this Appendix we thus present two publications I co-authored where some of that analysis was done.

The main topic of the publications is that of the representation of proteins as contact maps (equivalent to networks of residue interactions) and how one can reconstruct back the 3-dimensional structures from them. This is in fact connected to the main topic of this thesis since contact maps can also be used to represent the interaction between two different protein chains and thus can offer a neat simplified representation of a protein interface. This for instance has a very real connection to the current project as the interface calculation algorithm was based in the algorithm developed to calculate the intra-chain contacts (see the Appendix to Chapter 2).

The first publication "Optimal contact definition for reconstruction of contact maps" deals with the issue of determining what constitutes a good contact decomposition for a protein based on its ability to reconstruct its 3-dimensional structure. The second publication "CMView: interactive contact map visualization and analysis" is

an Application Note for the CMView software package developed at Michael Lappe's group. The software constitutes a very convenient and powerful analysis toolbox in dealing with proteins and contact maps, especially thanks to its connection to the well-known PyMOL molecular viewer which allows for immediate visualization of the contacts in the 3D structure.

Duarte *et al. BMC Bioinformatics* 2010, **11**:283
http://www.biomedcentral.com/1471-2105/11/283

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Optimal contact definition for reconstruction of Contact Maps

Jose M Duarte*[1,2], Rajagopal Sathyapriya[1], Henning Stehr[1], Ioannis Filippis[1,3] and Michael Lappe[1]

Abstract

Background: Contact maps have been extensively used as a simplified representation of protein structures. They capture most important features of a protein's fold, being preferred by a number of researchers for the description and study of protein structures. Inspired by the model's simplicity many groups have dedicated a considerable amount of effort towards contact prediction as a proxy for protein structure prediction. However a contact map's biological interest is subject to the availability of reliable methods for the 3-dimensional reconstruction of the structure.

Results: We use an implementation of the well-known distance geometry protocol to build realistic protein 3-dimensional models from contact maps, performing an extensive exploration of many of the parameters involved in the reconstruction process. We try to address the questions: a) to what accuracy does a contact map represent its corresponding 3D structure, b) what is the best contact map representation with regard to reconstructability and c) what is the effect of partial or inaccurate contact information on the 3D structure recovery. Our results suggest that contact maps derived from the application of a distance cutoff of 9 to 11Å around the C_β atoms constitute the most accurate representation of the 3D structure. The reconstruction process does not provide a single solution to the problem but rather an ensemble of conformations that are within 2Å RMSD of the crystal structure and with lower values for the pairwise average ensemble RMSD. Interestingly it is still possible to recover a structure with partial contact information, although wrong contacts can lead to dramatic loss in reconstruction fidelity.

Conclusions: Thus contact maps represent a valid approximation to the structures with an accuracy comparable to that of experimental methods. The optimal contact definitions constitute key guidelines for methods based on contact maps such as structure prediction through contacts and structural alignments based on maximum contact map overlap.

Background

For over 30 years [1,2] contact maps have been used as an alternative representation of protein structures. A contact map is a 2-dimensional representation of the residue interactions in a protein structure. This 2-dimensional representation takes the form of a binary matrix. A given cell (i, j) of the matrix can only take two values, 1 if the residues i and j are in contact or 0 otherwise. The definition of interaction varies but it is usually based on some cut-off distance between the atoms of the two residues. One can also see this description from another perspective as a residue interaction graph (RIG) with residues as nodes and the contacts as edges. In this view the binary matrix is no more than the adjacency matrix representing the graph.

Although they constitute a simple 2-dimensional representation of the molecule, contact maps still capture all important features of a protein fold. As such they are an invaluable tool for the analysis of biological macromolecules. They provide a computationally tractable representation of an otherwise complex problem, with the important advantage of being structural descriptors independent of the coordinate frame. Thus providing a sort of internal coordinates description, rotationally and translationally independent. However the simplified representation loses on accuracy as compared to the original 3-dimensional model. Multiple applications can be found in the literature that make use of the concept. Contact maps have been used for development of structural alignment algorithms [3,4], for automatic domain identification

* Correspondence: duarte@molgen.mpg.de
[1] Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany
Full list of author information is available at the end of the article

[5,6], in structural modelling by the extraction of contact-based empirical potentials [7-10] or for the identification of residues critical for folding [11], stability [12] and function [13]. Furthermore they have been used as a proxy for 3-dimensional structure prediction by means of machine learning techniques in order to predict residue contacts from sequence information [14-18].

Several methods have been proposed in the past for the reconstruction of contact maps. Most of them develop around the common mathematical theory of distance geometry first applied to chemistry by Blumenthal [19]. The theory took really off when Crippen and Havel [20] applied it to the problem of protein structure determination by NMR methods. In a typical NMR experiment distances between spatially close Hydrogen atoms can be determined for a protein in solution through the detection of the Nuclear Overhausser Effect (NOE) [21]. The NOE data can be seen then as a set of distance ranges between some pairs of Hydrogen atoms. Distance geometry deals with distances between points and their embedding in 3-dimensional space. In principle given a proper metric matrix with all exact distances among a set of points an analytical solution to the embedding can be found easily. The problem becomes more complicated when not all distances are given (sparse distance map) and when only distance ranges rather than exact distances are known. This is the case of the NMR experiments and equivalently of contact maps: we know some distance ranges between pairs of atoms for which we would like to find 3-dimensional coordinates. A heuristic algorithm (named EMBED) to solve the problem was proposed by Crippen and Havel and has been applied extensively ever since. Other algorithms have been proposed such as the alternating projection algorithm by Glunt et al. [22] or the geometric build-up algorithm by Wu and Wu [23].

However the problem of reconstructability of protein contact maps has not been fully addressed in the literature. A few studies [24-26] have tried to evaluate the accuracy of the existing methods but they all lack in completeness of the test set and thorough assessment of the different parameters or do not provide fully realistic protein models but only C_α traces.

Our aim here is twofold. We would like to find what is the reconstruction accuracy for an average protein so that the limits of the utility of contact maps in protein structure prediction can be precisely assessed. As a second aim we are looking for optimal criteria in the definition of a contact map decomposition model: atoms selected as interaction centres and distance cut-off. By decomposing a representative set of PDB protein structures into residue interaction graphs and then reconstructing them based purely on the contact information we should be able to assess the accuracy and loss of information in the decomposition process by comparing to the original native structure (see Figure 1). If a specific contact map model that reconstructs optimally can be found, that would help direct efforts in prediction of contact maps. Previous work has looked at optimality of contact definition from very different points of view, mainly in relation to how well contacting pairs describe the residue propensities when discriminating decoys from native structures. Here we look at it in a purely geometrical way, we are intending to find out how much of the 3D geometrical topology is captured by the network of contacts. Additionally by introducing artificial noise in the contact maps we also look at the effect of inaccurate contact information in the 3-dimensional recovery, essential to the applicability of contacts for predictive purposes.

Results and Discussion

We studied the reconstructability of a set of representative native PDB protein structures (see Methods). Firstly we decomposed the native proteins into contact maps with different contact type definitions and for several distance cut-offs. Then we used our reconstruction software to recreate the 3D structures based solely in the information supplied by the contact maps.

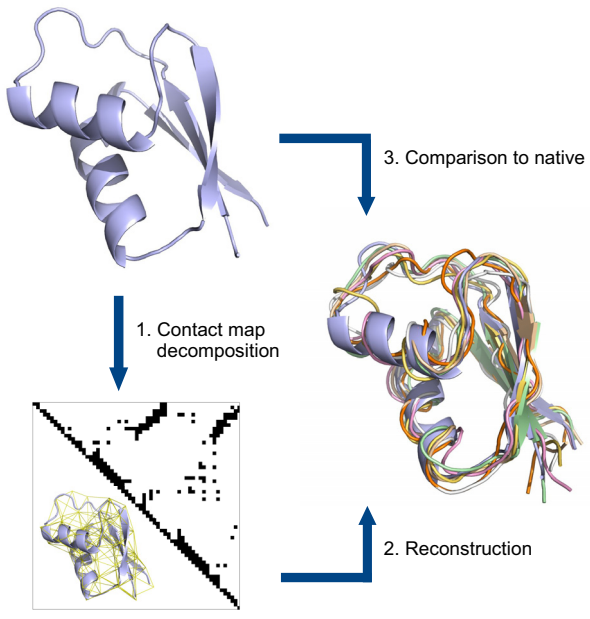


Figure 1 Schematic representation of the optimization procedure. 1) the native structure is decomposed into contact maps based on different definitions, 2) the 3D structure is reconstructed from contact information only, obtaining an ensemble of conformations, 3) the accuracy is measured against the original structure. The protein shown is PDB structure 1bxyA. The ensemble corresponds to 6 reconstructions (ribbon representation) in different colours and also contains the native protein (cartoon representation) in blue.

To measure the accuracy we then proceed by evaluating the RMSD of the generated models with the original structure. We measured the RMSD on the C_α atoms over all residues, independent of whether the reconstructions were based on C_α contact maps or not. This seems to be a well-established way of measuring the similarity between two structures especially when they are closely related and should facilitate the comparison to other published work. Another well-established method for structure comparison, GDT [27], was not deemed to be appropriate here as it is most useful in comparing structures over a broader range of dissimilarity as is the case in the CASP experiment.

Optimal cut-off

In Figure 2 we present the accuracy of reconstruction as measured by RMSD vs. the distance cut-off for contact maps based on C_α, C_β and C_α + C_β contact-types (see Methods for contact-type definitions).

The range of cut-offs chosen was based on values previously used in the literature keeping them within a biochemically sensible range: the minimum cut-off was 6Å as values below result in too sparse contact maps. At the other end we chose 15Å since beyond that the contact map starts to lose in information content becoming fully connected.

The first interesting observation is the existence of an optimal cut-off for all the contact types. This optimal value is not very precisely defined in most cases, it seems to span the cut-off distances from 9 to 11Å with higher cut-offs having only a marginal loss of accuracy. However we consider of a more significant value the lower cut-offs. First of all because of the biochemical meaning of the contacts. It is in the region about the 8Å cut-off where our definition of contact lead to distances between atoms that are in the range of the Van der Waals interactions. Also the information content of the contacts should be taken into account. As shown in Figure 3a the practically unchanged accuracy values in the higher cut-off regions are accompanied by an increase in the total number of contacts (the number of contacts increases roughly linearly with the distance cut-off). Thus we could see this as a loss of information content per contact i.e. we are adding a lot more information that is simply redundant. Figure 3b illustrates this better by representing the gain in accuracy with respect to contacts added vs the distance

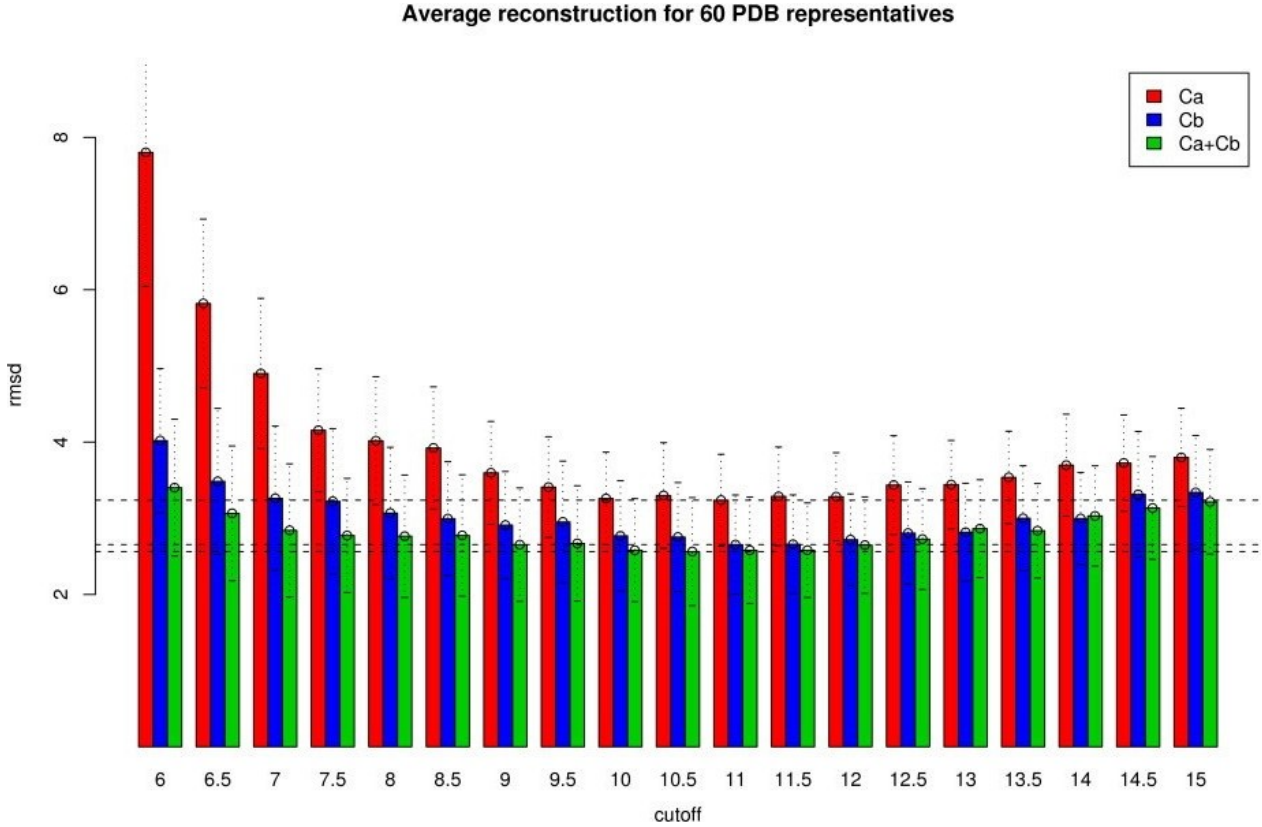


Figure 2 Accuracy of reconstructions. Reconstruction C_α RMSD vs. distance cutoff for each of the contact definitions. Plotted are the mean accuracy values for the set of 60 proteins for C_α, C_β and $C_\alpha + C_\beta$ contact definitions. Horizontal lines mark the minimum RMSD for each of them. The error bars represent the standard deviation across the distribution of 60 proteins.

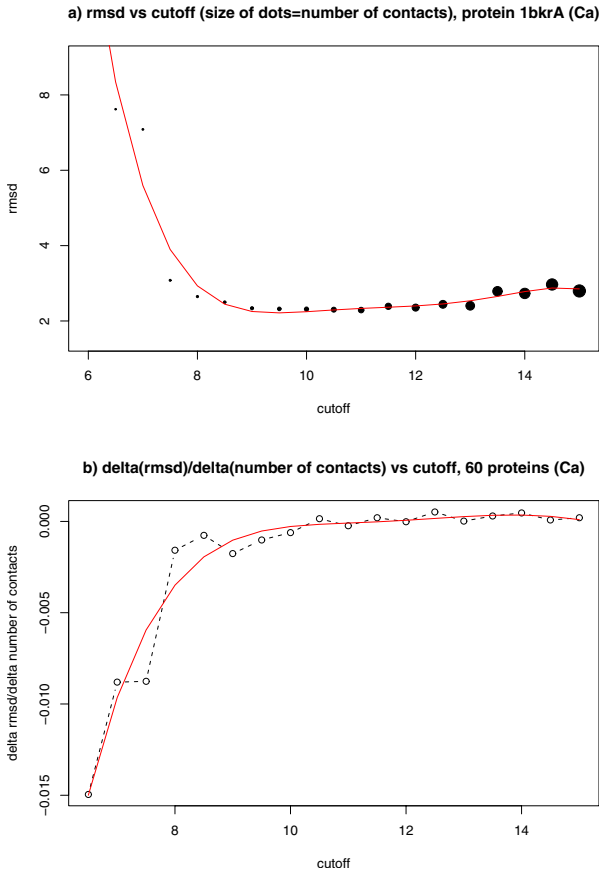


Figure 3 Number of contacts and reconstruction accuracy. a) RMSD values for the protein 1bkrA using C_α as contact definition, the size of the dots represent the total number of contacts in the contact map for a particular cutoff. The red curve is a linear fit to a polynomial. b) RMSD delta over delta of number of contacts against the cut-off for C_α contact definition for the average of the 60 proteins in the data set. The red curve is again a linear fit to a polynomial.

cut-off. The accuracy gain occurs only up to 8Å, after that there is no change as more contacts are added.

Additionally no dependence on the protein length across all cut-offs could be observed (see Figure 5). The reconstruction process seems to work with the same accuracy as measured by RMSD regardless of the protein size. This holds across all proteins tested (data not shown) and is in agreement with what similar studies found [26,24].

Our RMSD vs distance cut-off plots show no further improvement in accuracy beyond the optimal cut-off region. This is in clear disagreement with [26] where the reconstruction quality is reported to further increase for cut-off values as big as 18Å. This can be explained by the fundamentally different procedure of computing the reconstructed models: in our case an all atom approach with realistic regularization of the coordinates through a restraint-only harmonic potential was used for the construction of the models.

Vassura et al. on the other hand uses a simpler C_α trace model, without a final refinement phase. Optimal threshold values found here are in agreement to some of the reported optimal values found in other studies. There has been many attempts in the past to find an optimal contact map definition with respect to both distance cut-off and interaction centre. The optimizations were based in different criteria according to what the focus was in the particular study.

Some authors like Gromiha et al. [28] studied the correlation of relative contact order with folding rate, finding that from several cut-offs 8Å gave the best correlations for the C_α contact type when considering long range interactions only.

Karchin et al. [7] found that residue burial expressed as contact counts performs best at fold recognition for C_β contact type with a cut-off of 14Å. Similarly Benkert et al. [8] used the same residue burial measure and surprisingly found that a cut-off of 9Å was optimal, possibly due to differences in normalisation procedures. Quite a few studies tried to find an optimal contact definition based on the discriminatory power of contact-based empirical potentials in distinguishing decoys from native structures. Bolser et al. [9] found that the best performing two-body potential was that derived from C_β contact definition with a 12Å cut-off. Vendruscolo et al. [29] found that for the C_α contact type the best cut-off was at 8.5Å for a two-body contact potential.

As contact maps are only meaningful in the context of obtaining 3D protein models the reconstructability criterium should not be neglected when considering a contact definition for instance in the prediction of contacts. Contacts containing more geometrical information will be more valuable when building 3-dimensional models. This is of special importance if we consider that the reconstruction of contact maps seems to be possible even with sparser contact maps (see [30,31]), which means that contacts even at optimal definitions still seem to contain redundant information.

Optimal interaction centre

Comparing the accuracy values between the C_α, C_β and $C_\alpha + C_\beta$ cases (see Figure 2) it is apparent that $C_\alpha + C_\beta$ performs better across the whole range of cut-offs tested, with C_β alone doing also better than C_α. Figure 4 shows again this comparison for proteins divided into their respective SCOP classes. The trend holds within each of the SCOP classes.

Melo et al. [10] studying distance dependent empirical potentials explored several interaction centres concluding that the C_β atom was the best performing atom centre.

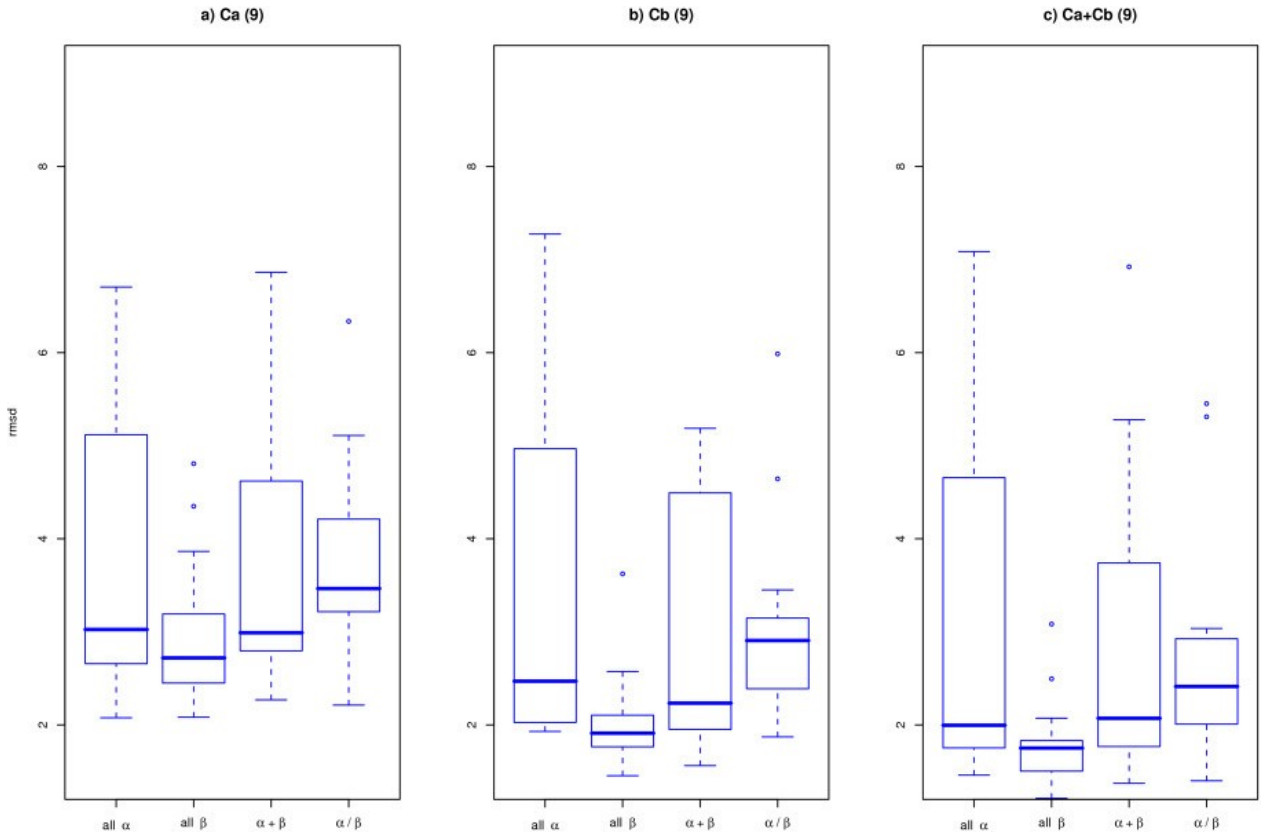


Figure 4 Variability for different SCOP classes. Reconstruction accuracy comparison for proteins in the four SCOP classes, using boxplots to depict the distributions of RMSD values. There are exactly 15 proteins per class from the set of 60 PDB representatives. a) For C_α b) for C_β and c) $C_\alpha + C_\beta$, all three at 9Å cutoff.

This seems to be a widely accepted result as indicates the use of the C_β contact type for the contact prediction category at the Critical Assessment of protein Structure Prediction (CASP) experiment [32].

Our study, purely based on the 3D geometrical information content of the contacts, confirms the preference for C_β as the interaction centre of choice. It seems natural that C_β is better in order to derive empirical potentials as it spans both the backbone and the side-chain. But also it is a superior point of choice for embedding a 3D structure from interatomic distance restraints. The interaction centre is able to capture geometrical information for the backbone positioning as well as for the orientation of the side-chain leading to a more precise 3D description.

Also of interest is the fact that the combination of both C_α and C_β contacts leads still to better reconstruction performance, indicating that there is some more backbone information not contained in the C_β restraints. This suggests an approach in the homology modelling of proteins based on distance restraints (see [33-35]): using two atoms per residue to restrain the geometry will lead to more precise models. We also obtained better accuracy results (data not shown) by choosing a backbone atom and a side-chain atom farther away from the C_β.

Reconstructions for different SCOP classes

We then address the question of whether the reconstruction process is dependant of the type of protein. In order to do so we separate our 60 proteins into the four SCOP classes to which they belong to, each of the classes containing 15 structures. Figure 4 shows the accuracy values for each of these four classes. The results hold for other cutoffs. It is striking that the accuracy and spread of the all-β group is significantly better than that of the other three. Interestingly the median values are not very far away for the 4 classes but the variances are hugely different especially for the all-*beta* case. Contrary to this result, in a similar study Saitoh et al. [24] stated that they did not encounter a dependency of the accuracy of reconstruction based on the SCOP class. This might be explained by the much smaller test set used in that study, 11 proteins in total and only 2 in the all-β class. Vassura et al. [26] did find some differences across different classes especially a lower accuracy for the all-*a* class, which we also observe here.

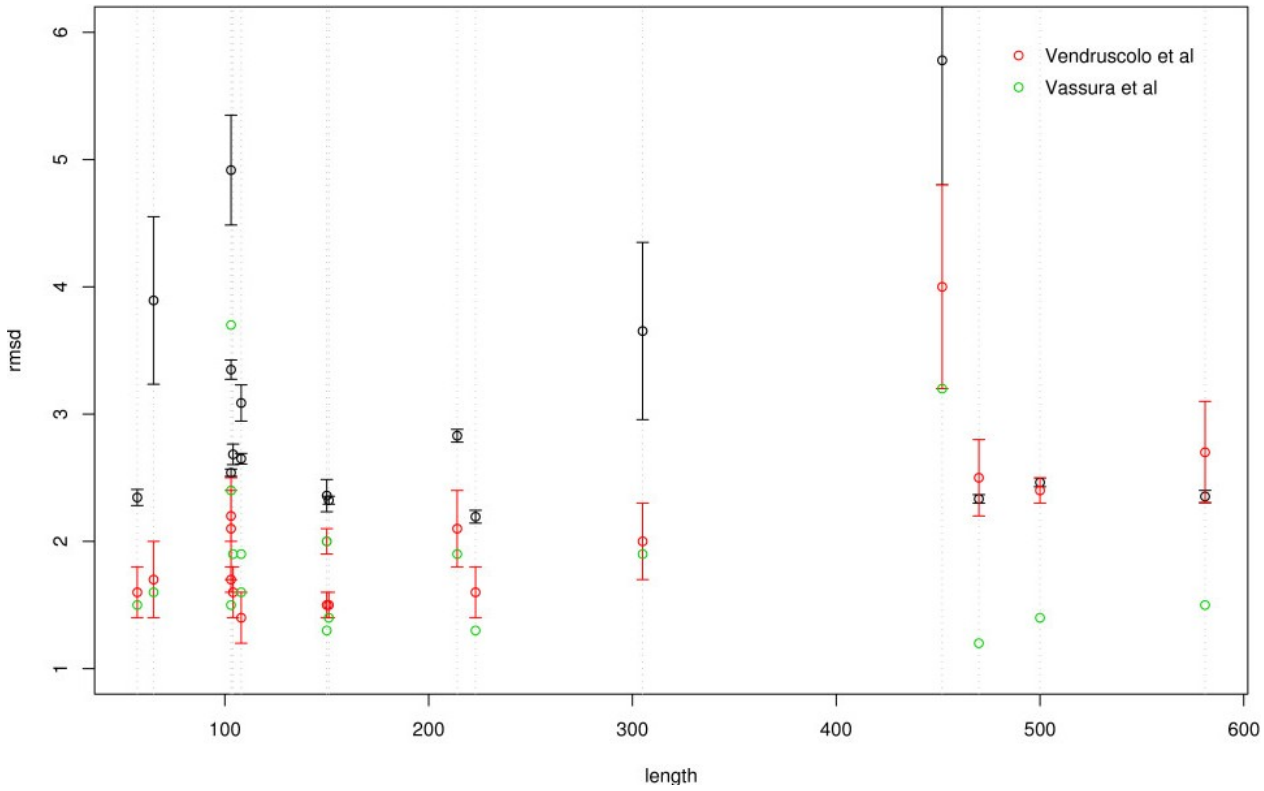


Figure 5 Comparison to previous studies. Comparison of our reconstruction RMSD values (black) with those of Vassura et al. (green) and Vendruscolo et al (red). The set is the one used by Vendruscolo and subsequently by Vassura. Two proteins were eliminated from their set because of ambiguities with the data. The error bars are for the variability across different runs (not reported by Vassura).

Variability of the reconstruction ensembles

The reconstruction process inherently leads to a non-unique solution fully matching the contact map. We studied the variance of the ensemble of reconstructed structures. The average spread of the pairwise RMSD among the ensemble structures is in most cases below 2Å. In Table 1 we present the spread values for a 12 proteins subset (see Methods). An example ensemble can be seen in Figure 1.

As seen in Figure 1 the reconstruction ensemble is reminiscent of an NMR structure ensemble, not surprisingly as both are based on fitting 3D coordinates to distance restraints. This shows another advantage of the contact map representation, namely that the conformational flexibility of the molecules is implicit in the model.

Comparison to previous studies

For completeness of this work we compare our results to those of two previously published reconstruction methods [26,25]. In Figure 5 we present our results (black) for the set of 17 proteins used by Vendruscolo et al. and subsequently by Vassura et al. together with their results (red and green respectively). Our RMSD values are higher in most cases. Remarkably the values of Vassura et al. are a lot lower. However caution should be taken in this comparison as they do not report on the variability (error) of the result. As their algorithm (like the others) is stochastic the evaluation of the variability across different runs is important to consider. Another important issue to take into account is that these two previous studies are using a simpler representation of proteins, namely one based on only the C_α atoms. In contrast here we are constructing full atom protein chains with realistic bonds and angles. This leads to higher RMSD values as more geometrical constraints need to be fulfilled.

Tolerance to missing contacts and noise

As a final part of the study we then address the question of reconstruction of contact maps in the more realistic scenario of incomplete or noisy maps, which is likely to be the case when the input is a predicted set of contacts. To do this instead of using real predictions, for instance from homology or machine learning methods, we simulate incomplete and noisy contact maps to thoroughly explore the effect of noise in the process of reconstruction.

Figure 6a presents the reconstruction accuracy versus the percentage of contact deletion. Thus we are simulating a prediction that misses contacts but with a 100% pre-

Table 1: RMSD of reconstruction ensembles.

PDB code	SCOP class	Length	Ensemble's average RMSD
1bkrA	all-α	109	1.93
1oddA	all-α	118	2.76
1cemA	all-α	363	1.69
1pzcA	all-β	123	1.52
1onlA	all-β	128	1.67
1eurA	all-β	365	2.49
1e6kA	α/β	130	1.91
1o8wA	α/β	146	1.71
1edeA	α/β	310	1.62
1r9hA	α + β	135	3.11
1ugmA	α + β	125	2.17
1iu4A	α + β	331	3.70

The 12 proteins subset with chain lengths and the average pairwise RMSD of the reconstruction ensembles, based on C_β contact maps with 8Å cut-off.

cision. The striking observation here is that the reconstruction seems to be very robust to missing information, thus indicating that there is a lot of redundancy in the contacts. A previous study in our group [30] deals with this problem in more depth and finds that one can even predict rationally a subset of contacts that somehow contain the most structural information.

Interestingly enough there seems to be a non-linear relationship in the information redundancy with respect to cut-off. Figure 6b represents as before the reconstruction RMSD versus the deletion of contacts but this time only for contact type C_β and different cut-offs. The loss of accuracy with lower percentage sampled subsets seems to decrease with higher cut-offs. Thus for the same percentage deletion one can recreate the original structure better with contact maps of higher cut-offs, i.e. the redundancy is higher. The second test that we perform intends to asses the robustness of the 3D recovery process with respect to the presence of noise, the case of a more realistic prediction with false positives. Figure 6c represents the reconstruction accuracy versus the percentage of noise added. The behaviour here is totally different than before. An addition of only 2% of random contacts severely affects the 3D recovery process. The C_β definition behaves better at all levels of noise.

An existing application [36] is reported to perform better with noisy contact maps, but this seems to be due to their pre-filtering based on finding well connected nodes, equivalent to finding contact clusters. As the test is against randomly added contacts this is not a very realistic filtering. In a real scenario a) one would not have all well-connected real contacts of the native map and b) the false positives would be very different from random noise. Thus we argue that the filtering used in FT-COMAR based in common neighbours is not realistic and so the reported tolerance to noise could not be extended to real situations. In our case we have tested the robustness of the algorithm still against random noise (which in principle would have a different distribution than predicted false positives) but we do not perform any pre-filtering. We believe this to constitute a more realistic benchmark.

The tests performed here are based on randomly generated inaccurate contact maps which in principle differ significantly from ab-initio predictions. However from our results here we could conclude that with adequately precise ab-initio contact predictions one could produce reasonable models. In fact we applied successfully some of these ideas in the CASP8 community-wide experiment for structure prediction [37]. In that case we used template-based contact maps that led to 3D models comparable to those of established methods. The non-random noise of the template-based maps did not seem to affect significantly the 3D recovery.

Conclusions

In this work we have studied the viability of computing 3D protein models from contact maps. We assessed the

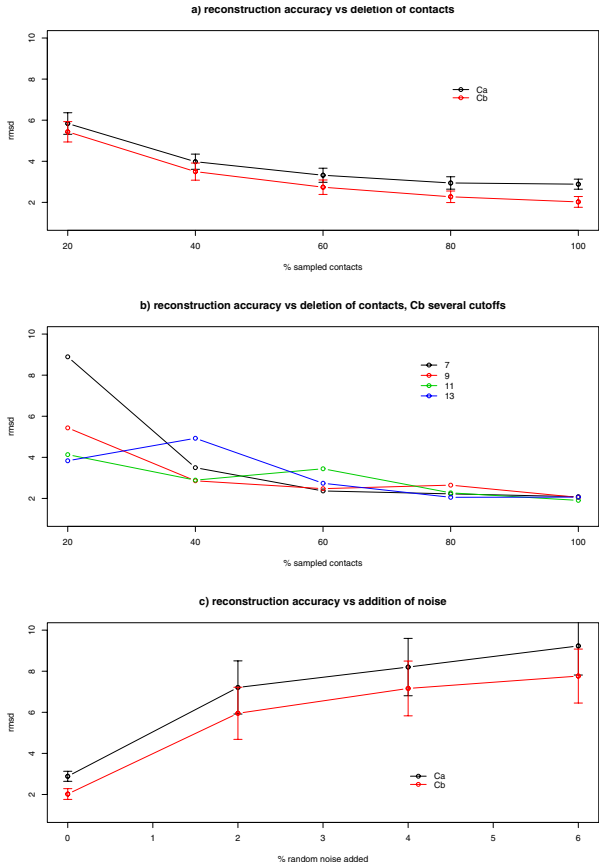


Figure 6 Reconstruction for incomplete or noisy maps. Behaviour of the reconstruction algorithm with noise or incomplete data. a) random subsets are sampled for C_α and C_β maps, b) random subsets are sampled for C_β maps at different cut-offs (7, 9, 11 and 13, with different colours) and c) random contact noise is added to the map (C_α and C_β maps). The 12 proteins subset (see Methods) was used for this analysis. For each of the levels of noise 10 random samples were taken and 30 models generated. The variability within the different proteins in the set is represented with the error bars.

performance of a reconstruction procedure based on the well known distance geometry protocol used extensively in NMR protein structure determination.

We perform a comprehensive evaluation covering a representative set of the PDB spanning the 4 SCOP classes. We then explore several possible contact map definitions and evaluate the accuracy of the reconstructions based on RMSD to the available native structure.

We found that contacts based on the C_β atoms are a better description of the 3-dimensional model than those based on C_α, confirming other studies that used one-body and two-body empirical contact-based potentials for fold recognition to find this optimum. Reconstruction accuracy can be further improved by using the two contact definitions together $C_\alpha + C_\beta$.

With regards to contact cut-offs we found that the optimal lies in the region from 9 to 11Å. We do not observe, contrary to previous studies [26] that the accuracy improves for higher cut-offs. Because of the increasing amount of contacts that higher cut-off contact maps yield, we preferred as an optimal threshold the lower end of the optimal range. A contact map based on a 9Å cut-off achieves maximal geometrical information per contact.

Interestingly the accuracy of the reconstruction seems to be different for different classes of proteins. Particularly the all-β SCOP class yields very good accuracies across all its members as compare to the other classes, leading to the conclusion that some topologies are more amenable to be described in terms of single atom distance restraints.

These results are particularly valuable for the contact prediction community. As contact prediction ultimately aims at obtaining 3-dimensional models of protein structures the usage of our optimal contact definition findings should contribute to better accuracies of the predictions. At the same time the results can be useful in the structural alignment of proteins through contact map overlap [3]. These methods seek a 3D alignment by optimising a contact map overlap measure. Clearly contacts that contain better 3-dimensional information should lead to improved results in the final alignments.

Further our 3D recovery procedure seems to perform also very well even if only a partial subset of the contacts is available. With as little as 40% of the contacts reasonably good models can be produced. On the contrary the method is very sensible to the presence of non-real contacts. The introduction of restraints at random points in the chain is simply fatal for the recovery of the original structure. This indicates that contact predictions should focus on accuracy rather than coverage.

Methods

Reconstruction pipeline

This study is based on the TINKER molecular dynamics package [38], available at http://dasher.wustl.edu/tinker. In particular the *distgeom* [39] program was used for the generation of 3-dimensional protein models from distance restraints which is at the core of the contact map reconstruction procedure.

An interface to the TINKER package was developed (Java) providing a single command line executable as a one stop solution for contact map reconstruction, taking contact maps as input and outputting PDB files. The software is multiplatform (Linux, Windows and Mac) and only requires a working copy of the TINKER package locally installed.

We have made our program freely available under the terms of the GPL v.2 at http://www.molgen.mpg.de/~lappe/reconstruct.

Reconstruction procedure

We generated distance restraints from the contact maps in the form of lower and upper bounds restraints for pairs of atoms (with standard value of 100.0 kcal/Å^2 for the force constant). The restraints were then fed into distgeom to generate a total of 30 models per structure using simulated annealing for refinement. The extensive study performed required a substantial amount of computation as we had 60 proteins, 3 contact-type definitions and 19 cutoff bins from 6 to 15 with 0.5 step. This gave a total of 3420 contact maps, for each of them we computed 30 structures in order to have a statistically meaningful sampling of the reconstruction space, resulting in a total of 102,600 models. The computations were carried out in a distributed fashion on a Linux cluster with over 100 CPUs.

The conformations found through the distance geometry protocol can not distinguish between the 2 enantiomers of the molecule, as chirality information is simply not present in the contact map. We overcome this problem by comparing to the native molecule through RMSD. The RMSD values for the conformation ensemble are found to be distributed bimodally, by simply choosing the lowest third of models as ranked by RMSD we are sure not to be falling into the wrong enantiomer.

Contact maps and distance restraints

We used two definitions of contact maps in this study: C_α and C_β. Two atoms were considered to constitute a contact when their euclidean distances where below the given cut-off. In the C_α model the backbone C_α atom for each residue is chosen, whilst for the C_β model the C_β atom of the side chain of each residue is taken, except for Glycine where we use the C_α atom.

For the reconstruction procedure we then need the contacts to be translated into distance restraints. Restraints were generated only for pairs of atoms corresponding to the contacts: C_α atoms or C_β atoms for each of the cases above. As upper bound of the restraint we used directly the distance cut-off, while for the lower bound value we used distance statistics derived from the PDB database. We proceeded by plotting the distance distribution for all C_α or C_β atoms and then choosing as our lower cutoff the value of the 90th percentile of the distribution.

Distance Geometry

The distance geometry procedure in TINKER is an implementation of the established distance geometry algorithms used for NMR protein structure determination, see [20]. Crippen and Havel proposed the EMBED algorithm consisting of three steps: bounds smoothing, embedding and regularization (coordinate refinement). The bounds smoothing is the procedure by which the initial sparse set of distance restraints is extended to obtain a full set of distance ranges for all pairs of atoms. This is achieved by means of the triangle inequality starting from the distances of known pairs. Once distance restraints are found for all pairs one only needs to select at random a particular value from within the restraints. There are several strategies for this selection [40], the most effective one is metrization. To perform metrization one proceeds starting at a random atom, choosing distances for it and then readjusting the whole matrix through the triangle inequality procedure. By doing this for all atoms the result is a sampled distance matrix where the triangle inequality is fulfilled or in other words a metric matrix. Once we have a distance matrix of exact distances for all pair of atoms a very good approximation of the 3-dimensional embedding can be obtained through the 3 largest eigenvalues of a certain transformation of the distance matrix. The result of the embedding is a good solution to the given distance restraints, however the geometry of the molecule is still not good enough especially with regards to the bond distances and angles. Thus the need for a final regularization step consisting in the minimization of an error function of the restraint violations usually done through simulated annealing.

Data set

In the selection of the data set we aimed at covering a diverse set of structures to ensure generality of the results obtained. We used a non-redundant PDB dataset of 60 proteins selected from SCOP release 1.73 [41]. Only monomeric, monodomain proteins from the four main SCOP classes and from highly populated folds are chosen. All proteins have resolutions better than 3.0Å, R-factor lower than 0.3 as well as no missing or ambiguous conformational data. A subset of 12 proteins, three per SCOP class, is selected from the dataset as used by Sathyapriya et al. [30]. From each group of 3 proteins, two fall in the size range of 100 - 120 amino acids and the third is three times as big as the other two. The PDB codes of the subset of proteins are given in Table 1.

Authors' contributions

JD performed the bulk of the analysis, developed the software and drafted the manuscript. RS performed the analysis related to error tolerance of reconstruction. HS developed the software and participated in the design of the study. IF selected the protein subset and contributed drafting the manuscript. ML initiated the study and participated in its design. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Dan Bolser for stimulating and fruitful discussions about the project.

Author Details

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany, [2]Laboratory of Biomolecular Research, Paul Scherer Institut, 5232 Villigen PSI, Switzerland and [3]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, UK

Received: 8 December 2009 Accepted: 27 May 2010
Published: 27 May 2010

References

1. Phillips DC: **The development of crystallographic enzymology.** *Biochem Soc Symp* 1970, **30:**11-28.
2. Nishikawa K, Ooi T, Isogai Y, Saito N: **Tertiary Structure of Proteins. I. Representation and Computation of the Conformations.** *Journal of the Physical Society of Japan* 1972, **32:**1331-1337.
3. Caprara A, Carr R, Istrail S, Lancia G, Walenz B: **1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap.** *J Comput Biol* 2004, **11:**27-52.
4. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.
5. Alexandrov N, Shindyalov I: **PDP: protein domain parser.** *Bioinformatics* 2003, **19(3):**429-430.
6. Emmert-Streib F, Mushegian A: **A topological algorithm for identification of structural domains of proteins.** *BMC Bioinformatics* 2007, **8:**237.
7. Karchin R, Cline M, Karplus K: **Evaluation of local structure alphabets based on residue burial.** *Proteins* 2004, **55(3):**508-518.
8. Benkert P, Tosatto SCE, Schomburg D: **QMEAN: A comprehensive scoring function for model quality assessment.** *Proteins* 2008, **71:**261-277.
9. Bolser DM, Filippis I, Stehr H, Duarte J, Lappe M: **Residue contact-count potentials are as effective as residue-residue contact-type potentials for ranking protein decoys.** *BMC Struct Biol* 2008, **8:**53.
10. Melo F, Sánchez R, Sali A: **Statistical potentials for fold assessment.** *Protein Sci* 2002, **11(2):**430-448.
11. Heringa J, Argos P: **Side-chain clusters in protein structures and their role in protein folding.** *J Mol Biol* 1991, **220:**151-171.
12. Capriotti E, Fariselli P, Rossi I, Casadio R: **A three-state prediction of single point mutations on protein stability changes.** *BMC Bioinformatics* 2008:S6.
13. Sathyapriya R, Vijayabaskar MS, Vishveshwara S: **Insights into protein-DNA interactions through structure network analysis.** *PLoS Comput Biol* 2008, **4(9):**e1000170.
14. Punta M, Rost B: **PROFcon: novel prediction of long-range contacts.** *Bioinformatics* 2005, **21(13):**2960-2968.
15. Cheng J, Baldi P: **Improved residue contact prediction using support vector machines and a large feature set.** *BMC Bioinformatics* 2007, **8:**113.
16. Pollastri G, Baldi P: **Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners.** *Bioinformatics* 2002:S62-S70.
17. Fariselli P, Casadio R: **A neural network based predictor of residue contacts in proteins.** *Protein Eng* 1999, **12:**15-21.
18. Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R: **The pros and cons of predicting protein contact maps.** *Methods Mol Biol* 2008, **413:**199-217.
19. Blumenthal LM: *Theory and Applications of Distance Geometry* Oxford University Press, Oxford; 1953.
20. Crippen GHHT: *Distance Geometry and Molecular Conformation* John Wiley & Sons, New York; 1988.
21. Wüthrich K: **Protein structure determination in solution by nuclear magnetic resonance spectroscopy.** *Science* 1989, **243(4887):**45-50.
22. Glunt W, Hayden TL, Raydan M: **Molecular conformations from distance matrices.** *J Comput Chem* 1993, **14:**114-120.
23. Wu D, Wu Z: **An updated geometric build-up algorithm for solving the molecular distance geometry problems with sparse distance data.** *J of Global Optimization* 2007, **37(4):**661-673.
24. Saitoh S, Nakai T, Nishikawa K: **A geometrical constraint approach for reproducing the native backbone conformation of a protein.** *Proteins* 1993, **15(2):**191-204.
25. Vendruscolo M, Kussell E, Domany E: **Recovery of protein structure from contact maps.** *Fold Des* 1997, **2(5):**295-306.
26. Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R: **Reconstruction of 3D structures from protein contact maps.** *IEEE/ACM Trans Comput Biol Bioinform* 2008, **5(3):**357-367.
27. Zemla A, Venclovas C, Moult J, Fidelis K: **Processing and analysis of CASP3 protein structure predictions.** *Proteins* 1999:22-29.
28. Gromiha MM, Selvaraj S: **Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction.** *J Mol Biol* 2001, **310:**27-32.
29. Vendruscolo M, Najmanovich R, Domany E: **Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading?** *Proteins* 2000, **38(2):**134-148.
30. Sathyapriya R, Duarte JM, Stehr H, Filippis I, Lappe M: **Defining an Essence of Structure Determining Residue Contacts in Proteins.** *PLoS Comput Biol* 2009, **5(12):**e1000584.
31. Chen Y, Ding F, Dokholyan NV: **Fidelity of the protein structure reconstruction from inter-residue proximity constraints.** *J Phys Chem B* 2007, **111(25):**7432-7438.
32. Graña O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A: **CASP6 assessment of contact prediction.** *Proteins* 2005:214-224.
33. Havel TF, Snow ME: **A new method for building protein conformations from sequence alignments with homologues of known structure.** *J Mol Biol* 1991, **217:**1-7.
34. Aszódi A, Taylor WR: **Homology modelling by distance geometry.** *Fold Des* 1996, **1(5):**325-334.
35. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J Mol Biol* 1993, **234(3):**779-815.
36. Vassura M, Margara L, Lena PD, Medri F, Fariselli P, Casadio R: **FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps.** *Bioinformatics* 2008, **24(10):**1313-1315.
37. Stehr B, Duarte J, Filippis I, Rajagopal S, Syal K, Risbud S, Holm L, Lappe M: **StruPPi: comparative modeling using consensus information from multiple templates and physics-based refinement.** *Abstracts book, 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction* 2008.
38. Ponder JW: *TINKER - Software tools for molecular design, User's Guide for Version 4.2* Washington University School of Medicine; 2004.
39. Hodsdon ME, Ponder JW, Cistola DP: **The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm.** *J Mol Biol* 1996, **264(3):**585-602.
40. Kuszewski J, Nilges M, Brnger AT: **Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm.** *J Biomol NMR* 1992, **2:**33-56.
41. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247(4):**536-540.

doi: 10.1186/1471-2105-11-283
Cite this article as: Duarte *et al.*, Optimal contact definition for reconstruction of Contact Maps *BMC Bioinformatics* 2010, **11**:283

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 11 2011, pages 1573–1574
doi:10.1093/bioinformatics/btr163

Structural bioinformatics Advance Access publication April 5, 2011

CMView: Interactive contact map visualization and analysis

Corinna Vehlow[1], Henning Stehr[2], Matthias Winkelmann[2], José M. Duarte[3], Lars Petzold[2], Juliane Dinse[1] and Michael Lappe[2,*]

[1]Department of Simulations and Graphics, O.v.G. University Magdeburg, [2]Max Planck Insitute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany and [3]Laboratory of Biomolecular Research, Paul Scherrer Institut, 5232 Villigen PSI, Switzerland

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Contact maps are a valuable visualization tool in structural biology. They are a convenient way to display proteins in two dimensions and to quickly identify structural features such as domain architecture, secondary structure and contact clusters. We developed a tool called CMView which integrates rich contact map analysis with 3D visualization using PyMol. Our tool provides functions for contact map calculation from structure, basic editing, visualization in contact map and 3D space and structural comparison with different built-in alignment methods. A unique feature is the interactive refinement of structural alignments based on user selected substructures.

Availability: CMView is freely available for Linux, Windows and MacOS. The software and a comprehensive manual can be downloaded from http://www.bioinformatics.org/cmview/. The source code is licensed under the GNU General Public License.

Contact: lappe@molgen.mpg.de, stehr@molgen.mpg.de

Received on December 22, 2010; revised on March 2, 2011; accepted on March 27, 2011

1 INTRODUCTION

The tertiary structure of a protein is determined by non-covalent residue interactions. An all-atom distance map is a lossless representation of the 3D coordinates (save chirality). Distance data can be further reduced to a binary residue contact map while still allowing complete reconstructions within 2Å RMSD (Duarte *et al.*, 2010). The native fold is retained in reconstructions using sparse subsets comprising as low as 10% of native contacts (Sathyapriya *et al.*, 2009). Contact maps are a convenient way to highlight structural features like domain architecture, secondary structure and contact clusters and they display unique information about the sequence separation of contacting residues which is not easily visible in 3D representations. The study of contact maps has been a valuable source of insight in experimental and computational protein structure analysis. They have for example been used to measure the dissimilarity of structures (Caprara *et al.*, 2004), to analyze protein–protein interaction patterns (de Melo *et al.*, 2007) and to study protein folding (Vendruscolo and Domany, 2000). Here we present a tool which combines the strengths of contact map and 3D visualization for protein analysis.

*To whom correspondence should be addressed.

CMView implements several new features while seamlessly combining the functionality of different contact map visualization programs already developed. Among those, *Structer+Dotter* (Sonnhammer and Wooton, 1998) contains modules for the generation (*Structer*) and simple visualization (*Dotter*) of distance maps and contact maps from PDB files. *SeqX* (Biro and Fordos, 2005) integrates frequency counts of residue combinations. With *Protmap2D* (Pietal *et al.*, 2007) and *Con-StructMap* (Chung *et al.*, 2007) contact maps of two conformations can be compared side-by-side. *Con-Struct Map* also allows to compare non-sequence-identical proteins by loading an alignment from a file.

2 CMVIEW FEATURES

CMView is a stand-alone Java application for interactive visualization, analysis and manipulation of protein contact maps. It integrates protein analysis in contact map and 3D space via an interface to the molecular viewer PyMol (DeLano, 2002). CMView is open source software licensed under the GNU General Public License (GPL). It is available for Mac OS X, Linux, Windows and other platforms supporting Java 6.

Contact information can be read from various sources and file formats (PDB, CASP TS, CASP RR, native CSV) either from local files or directly from the PDB website. The contact definition can be specified in terms of contact type (all-atom, C-α, C-β) and contact threshold (distance cutoff in Å). The main application window (Fig. 1, right) shows the contact map and the various menu options for editing and analysis. If 3D coordinates are provided, the structure will be shown in a separate PyMol window (Fig. 1, left), making the full set of advanced visualization features of PyMol available. Any selection of (several) contacts can be exported as a corresponding PyMOL selection. In addition, the two views are intimately linked: the current residue pair underneath the crosshair in the 2D map is continuously displayed as an edge annotated with the euclidian distance in the 3D scene. The distance map, contact density and triangle inequality relations (here called common neighborhoods) are available as colored overlays in the contact map window.

2.1 Selecting, editing and export

CMView offers various functions to manipulate the contact map. Contacts can be selected individually or by using rectangle select, fill select, diagonal select and neighbor select tools that resemble similar functions known from graphics applications. Entire sets of contacts can be deleted or highlighted in different colors and exported to the

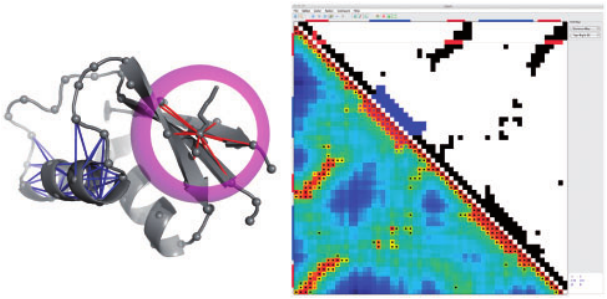


Fig. 1. Screenshot of *CMView*: 3D structure (grey backbone, left), contact map (C-α 8Å, upper right) and distance map (lower right) of Ribosomal Protein L30 (PDB code 1bxyA). Contacts within the n-terminal α-helix and subsequent turn are shown in blue. The contact sphere denoting the distance cutoff (here 8Å) is displayed for residue LYS5 along with its contacts in red. C-α positions are indicated as grey spheres.

3D structure. The option to toggle individual contacts gives full user control in editing the map.

As an additional feature, *CMView* implements the Cone-Peeling Algorithm (Sathyapriya *et al.*, 2009), which computes a fast approximation of an *essential* subset of contacts sufficient to maintain the native fold when reconstructing the 3D structure using distance geometry. Export options include text files in CASP RR or native CSV format for further processing. The map visualization, including highlighted contacts, can be exported as a PNG image.

2.2 Pairwise comparison

A key feature of *CMView* is the pairwise structural comparison of two proteins or conformations. For this purpose, a second structure can be loaded and aligned via one of the following methods: Needleman-Wunsch sequence alignment (Moustafa, 2007), SADP contact-based structural alignment (Jain and Lappe, 2007) or Dali structural alignment (Holm and Sander, 1995). The comparison view allows quick identification of shared and unique contacts. Common contacts are shown in black and contacts that are unique to one structure are shown in pink (for the first structure, e.g. a predicted structure) and green (for the second structure, e.g. the native structure). The two structures are also superimposed in the PyMol window, doing a best fit on the residues that are in contact in both structures. The 3D alignment can be interactively refined by selecting contacts and recalculating the superposition based on this subset. This feature allows the comparison of different alignments based on shared substructures in cases where a global rigid-body alignment is not optimal. To our knowledge, CMView is the only application that allows such an alignment of substructures in an interactive fashion.

3 CONCLUSION

CMView combines the strengths of rich contact map analysis with traditional 3D visualization in a single application. As a tool for contact map generation, modification and analysis it is the most feature complete application to date. Special emphasis has been put on integrating tools for the analysis of secondary structure interaction patterns and for the pairwise comparison of structural models or related proteins. In particular, structural alignments can be interactively refined based on different subsets of shared contacts. The real-time link of the current position in the 2D contact map with highlighting the corresponding residue pair in the structure provides a combined '2D/3D-cursor'. This offers a more intuitive approach to explore the relationships between contact patterns and protein structure. These unique features make CMView a valuable tool for structural analysis, protein modeling, assessment of structure predictions and education in structural biology.

4 FUTURE DIRECTIONS

Future work will include full integration of a reconstruction engine to transform modified contact maps back into 3D structures. This will allow CMView to be used as an interactive, contact-based protein modeling tool. An experimental command line tool called 'reconstruct' which allows such 3D reconstructions via an interface to a distance geometry package is already available for download from the CMView website. Preview snapshots of future versions will be available from the authors upon request.

ACKNOWLEDGEMENTS

We thank Brijnesh J. Jain, Dan M. Bolser and Ioannis Filippis for helpful comments and suggestions.

Funding: Max Planck Society.

Conflict of Interest: none declared.

REFERENCES

Biro,J.C. and Fordos,G. (2005) Seqx: a tool to detect, analyze and visualize residue co-locations in protein and nucleic acid structures. *BMC Bioinformatics*, **6**, 170.

Caprara,A. *et al.* (2004) 1001 optimal PDB structure alignments: integer programming methods for finding the maximum contact map overlap. *J. Comput. Biol.*, **11**, 27–52.

Chung,J.L. *et al.* (2007) Con-struct map: a comparative contact map analysis tool. *Bioinformatics*, **23**, 2491–2492.

DeLano,W. (2002) *The PyMOL Molecular Graphics System*. Available at http://pymol.sourceforge.net/.

de Melo,R. *et al.* (2007) Finding protein-protein interaction patterns by contact map matching. *Genet. Mol. Res.*, **6**, 946–963.

Duarte,J.M. *et al.* (2010) Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, **11**, 283–283.

Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.

Jain,B.J. and Lappe,M. (2007) *Joining Softassign and Dynamic Programming for the Contact Map Overlap Problem*, vol. 4414 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 410–423.

Moustafa,A. (2007) Jaligner: Open source java implementation of Smith-Waterman. Availabe at http://jaligner.sourceforge.net.

Pietal,M.J. *et al.* (2007) Protmap2d: visualization, comparison and analysis of 2d maps of protein structure. *Bioinformatics*, **23**, 1429–1430.

Sathyapriya,R. *et al.* (2009) Defining an essence of structure determining residue contacts in proteins. *PLoS Comput. Biol.*, **5**, e1000584.

Sonnhammer,E.L.L. and Wooton,J.C. (1998) Dynamic contact maps of protein structures. *J. Mol. Graph. Model.*, **16**, 1–5.

Vendruscolo,M. and Domany,E. (2000) Protein folding using contact maps. *Vitam Horm*, **58**, 171–212.

Acknowledgements

Many people participated directly or indirectly in the work presented. My thanks go to all these people:

Guido Capitani for offering me the possibility to work in his team and for being a fantastic supervisor always ready to listen to my questions about the intricate details of biochemistry and biology.

My faculty responsible at the University of Zurich, Amedeo Caflisch for so kindly agreeing to head my thesis committee. Also to Raimund Dutzler for being member of the thesis committee.

The rest of the Capitani team: Martin Schärer, Nikhil Biyani and Kumaran Baskaran for all the hours shared hacking along in OFLC/110.

Adam Srebniak from SyBIT who did a herculean job in implementing a sophisticated web interface for EPPIC. Also to the head of SyBIT Peter Kunszt, who supported us with infrastructure and technical help in our complicated IT demands.

The scientific IT team at PSI: Derek Feichtinger and Valeri Markushin for the support in setting up the server infrastructure for EPPIC and for running Merlin4 (the PSI cluster) so smoothly. Peter Hüsser also from the PSI IT department for setting up and taking care of the Virtual Machine infrastructure and of the new EPPIC hardware. Also an important thanks goes to the security and network people for hearing my continuous complaints about network security.

The rest of the LBR people which made my time at PSI so enjoyable and interesting. Special thanks to the "lunch club": Sophie Demarche, Cristina Manatschal, Manuel Hilbert, Ingrid Imhof, Bara Malkova and many many others who made all those lunch and coffee breaks so enjoyable.

Gebhard Schertler, head of the LBR, for making the lab such a dynamic and interesting place to work in.

My former colleagues at the Max Planck Institute for Molecular Genetics in Berlin: Michael Lappe, Henning Stehr, Dan Bolser, Ioannis Filippis, Sathyapriya Rajagopal, Matthias Winkelmann, Lars Petzold, Juliane Dinse, Corinna Vehlow and Bosco Ho. I have to thank them most of all for introducing me to the structural bioinformatics world. Many of them also contributed directly to the project by writing code for the OWL library and the CMView contact map viewer. Special thanks go to Henning Stehr for as well translating the thesis summary into German. Also to Bosco Ho, who

I could not meet yet but who has contributed implementations of a few algorithms in the OWL code.

Many thanks as well to all the anonymous reviewers that spent time in reviewing the EPPIC paper and for all their great suggestions. Also to the organizers of the UniProt10 symposium and the 3DSIG meeting for giving me the opportunity to talk about my work.

And last but not least to my friends and family. Special mention to my wife Mayra Nevarez not only for being always patient and supportive but also for her very direct contributions in creating the EPPIC logo, helping designing the EPPIC web site and creating some of the figures in this thesis.

I would also like to gratefully acknowledge funding to the project from the PSI Forschungskommission, SyBIT and the Swiss National Science Foundation (SNF).

Curriculum Vitae

Jose Manuel DUARTE
Born 25.04.1975 in Cadiz, Spain
Spanish nationality

Education

2010 – 2013 **PhD in Structural Bioinformatics.** University of Zurich and Laboratory of Biomolecular Research, Paul Scherrer Institute (Villigen, Switzerland). Supervision by Dr. Guido Capitani.

2001 – 2003 **MSc in Bioinformatics.** Birkbeck College, University of London (London, UK). Thesis title: "Error analysis in Oligonucleotide Microarrays"

1993 – 1999 **MSc in Physics (Licenciatura en Fisica).** Universidad Complutense de Madrid (Madrid, Spain). Material Science specialization.

1989 – 1993 **High School Education (Bachillerato).** Instituto de Enseñanza Secundaria Pedro Muñoz Seca (El Puerto de Santa Maria, Cadiz, Spain)

Relevant Professional Experience

2005 – 2010 **Bioinformatics Software Developer.** Max Planck Institute for Molecular Genetics (Berlin, Germany), Dr. Michael Lappe's group.

2002 – 2005 **Bioinformatics Systems Administrator and Analyst for Microarray Facility.** MRC Functional Genetics Unit, University of Oxford (Oxford, UK), Prof. Chris Ponting's group.

Publications

Protein interface classification by evolutionary analysis. **Jose M Duarte**, Adam Srebniak, Martin A Schärer and Guido Capitani, BMC Bioinformatics, 2012, 13:334. PMID: 23259833

The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Henning Stehr, Seon-Hi Jang, **Jose M Duarte**, Christoph Wierling, Hans

Lehrach, Michael Lappe, Bodo MH Lange, Molecular Cancer, 2011, 10:54. PMID: 21575214

CMView: interactive contact map visualization and analysis. Corinna Vehlow, Henning Stehr, Matthias Winkelmann, **Jose M Duarte**, Lars Petzold, Juliane Dinse and Michael Lappe. Bioinformatics, 2011, 27(11):1573-4. PMID: 21471016

Optimal contact definition for reconstruction of Contact Maps. **Jose M Duarte**, Rajagopal Sathyapriya, Henning Stehr, Ioannis Filippis and Michael Lappe. BMC Bioinformatics, 2010, 11:283. PMID: 20507547

PDBWiki: added value through community annotation of the Protein Data Bank. Henning Stehr, **Jose M Duarte**, Michael Lappe, Jong Bhak, Dan M Bolser. Database: The Journal of Biological Databases and Curation, 2010; doi: 10.1093/database/baq009. PMID: 20624717

Defining an Essence of Structure Determining Residue Contacts in Proteins. Rajagopal Sathyapriya, **Jose M Duarte**, Henning Stehr, Ioannis Filippis, Michael Lappe, PLoS Computational Biology, 2009, 5(12): e1000584, PMID: 19997489

Designing evolvable libraries using multi-body potentials. Michael Lappe, Ganesh Bagler, Ioannis Filippis, Henning Stehr, **Jose M Duarte**, Rajagopal Sathyapriya, Current Opinion in Biotechnology, 2009, PMID: 19713097

Residue contact-count potentials are as effective as residue-residue contact-type potentials for ranking protein decoys. Dan Bolser, Ioannis Filippis, **Jose M Duarte**, Henning Stehr, Michael Lappe, BMC Structural Biology, 2008 Dec, PMID: 19063740

Gene expression profiling studies on Caenorhabditis elegans dystrophin mutants dys-1(cx-35) and dys-1(cx18). Paula R Towers, Pascal Lescure, Dilair Baban, Julie A Malek, **Jose Duarte**, Emma Jones, Kay E Davies, Laurent Ségalat and David B Sattelle, Genomics, 2006 Nov;88(5):642-9, PMID: 16962739

Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Rat Genome Sequencing Consortium, Nature. 2004 Apr 1;428(6982):493-521, PMID: 15057822

Oral presentations

8th Sept 2012 *The footprint of evolution on protein-protein interfaces through UniProt history.* Symposium: The first 10 years of UniProt. Basel, Switzerland.

19th Jul 2013 *Sequence comes to the structural rescue: identifying relevant protein interfaces in crystal structures.* 3DSIG Satellite Meeting of ISMB 2013 conference. Berlin, Germany.

Courses and Meetings

5th-9th Dec 2010 Poster presentation at the 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP9). Asilomar, California, USA

8th-13th Jan 2012 10th NCCR practical course: Introduction to Biomolecular Modelling. Kandersteg, Switzerland.

9th-12th Sep 2012 Poster presentation at the European Conference in Computational Biology, ECCB 2012. Basel, Switzerland.

19th-23rd Jul 2013 21st International Conference on Intelligent Systems for Molecular Biology ISMB/ECCB 2013. Berlin, Germany.